Frustratingly Easy Truth Discovery

Reshef Meir[1], Ofra Amir[1], Omer Ben-Porat[1], Tsviel Ben-Shabat[1], Gal Cohensius[1], Lirong Xia[2]

[1] Technion—Israel Institute of Technology
[2] Rensselaer Polytechnic Institute (RPI)
{reshefm, oamir, omerbp}@ie.technion.ac.il, {tsviel,galcohensius}@gmail.com, xial@cs.rpi.edu

Abstract

Truth discovery is a general name for a broad range of statistical methods aimed to extract the correct answers to questions, based on multiple answers coming from noisy sources. For example, workers in a crowdsourcing platform. In this paper, we consider an extremely simple heuristic for estimating workers' competence using average proximity to other workers. We prove that this estimates well the actual competence level and enables separating high and low quality workers in a wide spectrum of domains and statistical models. Under Gaussian noise, this simple estimate is the unique solution to the Maximum Likelihood Estimator with a constant regularization factor.

Finally, weighing workers according to their average proximity in a crowdsourcing setting, results in substantial improvement over unweighted aggregation and other truth discovery algorithms in practice.

Introduction

> "All happy families are alike; each unhappy family is unhappy in its own way."
>
> — Leo Tolstoy, Anna Karenina

Consider a standard crowdsourcing task such as identifying which images contain a person or a car (Deng et al. 2014), or identifying the location in which pictures were taken (McLaughlin 2014). Such tasks are also used to construct large datasets that can later be used to train and test machine learning algorithms. Crowdsourcing workers are usually not experts, thus answers obtained this way often contain many mistakes (Vuurens, de Vries, and Eickhoff 2011; Wais et al. 2010), and multiple answers are aggregated to improve accuracy.

From a theory/statistics perspective, "truth discovery" is a general name for a broad range of methods that aim to extract some underlying ground truth from noisy answers. While the mathematics of truth discovery dates back to the early days of statistics, at least to the *Condorcet Jury Theorem* (Condorcet 1785), the rise of crowdsourcing platforms suggests an exciting modern application of aggregating *complex labels* from varied domains such as image processing and natural language, to healthcare. For example, the Etch-a-Cell project uses volunteers to trace the boundary of tumors on Electron Microscopy images ((Spiers et al. 2021), see Fig. 1).

Yet, the vast majority of the theoretical literature on truth discovery follows Condorcet by focusing on binary, multi-label or sometimes real-valued questions (see Related Work section), while specific applications with complex labels often rely on specialized algorithms.

Many of these algorithms aim to identify first the most competent workers. While some of them employ highly sophisticated analysis, others are much more direct: for example, Kobayashi (2018) suggests a 'frustratingly easy' algorithm that ranks workers by their *average cosine similarity* to others in a text summarization task; and Kurvers et al. (2019) prove that the *Hamming distance* of a worker from others is correlated with her competence in answering yes/no questions. Of course, using average similarity or distance is not a new idea, and is extensively employed outside the context of aggregation, for example in *Games with a Purpose* (Von Ahn and Dabbish 2008; Huang and Fu 2013) to identify outliers, and in peer prediction to incentivize effort (Witkowski et al. 2013).

In this paper we argue that average similarity is a powerful tool, with nothing special about Cosine or Hamming similarity in particular. Our main observation can be written as follows:

Theorem (Anna Karenina principle, informal)**.** *The expected average similarity of each worker to all others, is roughly linearly increasing in her competence.*

Essentially, the theorem says that as in Tolstoy's novel, "good workers are all alike," whereas "each bad worker is bad in her own way" and thus not similar to other workers.

Contribution and Paper Structure

After the preliminary definitions , we prove a formal version of the Anna Karenina principle and show how it can be used to identify poor workers without assuming specific label structure. We show how additional assumptions lead to tighter corollaries of exactly or approximately linear relation between pairwise similarity and competence. To the best of our knowledge these are the first formal guarantees on general-domain truth discovery.

In Section the next section we focus on the widely studied case of Gaussian noise. We prove that the average distance to other workers coincides with the Maximum Likelihood Estimator (MLE) for workers' (in)competence—the first guarantee of this type regarding average similarity or distance.

We then explain how to leverage the Anna Karenina principle for aggregation using a simple algorithm (P-TD). We demonstrate on real and synthetic data, that P-TD substantially improves aggregation accuracy, competing well with advanced and domain-specific algorithms.

Most proofs, as well as additional empirical results are available in the full version of the paper on arXiv: https://arxiv.org/abs/1905.00629.

Related Work

The Condorcet Jury Theorem (Condorcet 1785) was perhaps the first formal treatment of truth discovery, and extensions to experts with heterogeneous competence levels were surveyed by Grofman, Owen, and Feld (1983). The idea of estimating workers' competence in order to improve aggregation is thus underlying many of the algorithms in the area (a recent survey is in (Li et al. 2016)). We should note that *self-reporting* of accuracy often leads to poor results (Gadiraju et al. 2017; Prelec, Seung, and McCoy 2017).

Average similarity We have mentioned in the introduction the two applications of average similarity to truth discovery that we are aware of. Both of them assume a specific label structure and (somewhat surprisingly) both are quite recent: Kobayashi (2018) proved that cosine similarity approximates a known kernel density estimator. Kurvers et al. (2019) focused on *binary questions with independent errors*, showing both theoretically and empirically that the expected average *Hamming proximity* correlates with the true competence, albeit without comparing to any other algorithm.

Our Anna Karenina theorem entails the Kurvers et al. result as a special case, and provides explicit performance guarantees for the heuristic suggested by Kobayashi.

Domain-specific algorithms Many truth-discovery algorithms have been proposed for specific label structures, mostly for categorical (multiple-choice) and real-valued labels. Often these algorithms entwine accuracy and ground truth estimation, by iteratively aggregating labels to obtain an estimate of the ground truth, and using that in turn to estimate workers competence. This approach was pioneered by the EM-style Dawid-Skene estimator (Dawid and Skene 1979), with many follow-ups (Karger, Oh, and Shah 2011; Gao and Zhou 2013; Aydin et al. 2014; Xiao et al. 2016; Zhao and Han 2012; Li et al. 2012).

Another class of algorithms uses spectral methods to infer the competence and/or other latent variables from the covariance matrix of the workers (Parisi et al. 2014; Zhang et al. 2016), or from their pairwise Hamming similarity (Li, Baba, and Kashima 2018). Note that covariance can also be thought of as a measure worker similarity in the context of binary labels. In rank aggregation, every voting rule can be considered as a truth-discovery algorithm (Mao, Procaccia, and Chen 2013; Caragiannis, Procaccia, and Shah 2013).

Some of these works also provide formal convergence guarantees and/or bounds on the error that are subject to assumptions on the distribution of answers.

General labels When there are complex labels that are not numbers or categories, but for example contain text, graphics and/or hierarchical structure, there may not be a natural way to aggregate them but we would still want to evaluate workers' competence.

Two recent papers suggest to use the pairwise distance (or similarity) matrix as a general domain-independent abstraction, then applying sophisticated algorithms on this matrix: The *multidimensional annotation scaling* (MAS) model (Braylan and Lease 2020) extends the Dawid-Skene model by calculating the labels and competence levels that would maximize the likelihood of the observed distance matrix, using the Stan probabilistic programming language; Another approach is to find a 'core' of good workers (Kawase, Kuroki, and Miyauchi 2019), by looking for a dense subgraph of the similarity matrix.

While we adopt the approach that *pairwise similarity is the right domain-independent abstraction* for general labels, we argue that usually there is no need for such complex algorithms: a 'frustratingly easy' average is sufficient.

Preliminaries

We consider a set N of n workers, each providing a report in some space Z. We denote elements of Z (typically m-length vectors, see below) in **bold**. Thus, an instance of a truth discovery is a pair $\langle S = (\boldsymbol{s}_i)_{i\in N}, \boldsymbol{z}\rangle$, where $\boldsymbol{s}_i \in Z$ is the report of worker i, and $\boldsymbol{z} \in Z$ is the *ground truth*. S is also called a *dataset*.[1]

Noise model We do not make any assumptions regarding the ground truth $\boldsymbol{z}$. The *type* t_i of a worker determines her distribution of answers. A dataset is constructed in two steps:

(1) Sample a finite *population* of workers i.i.d from a distribution $\mathcal{T}$ (called a *proto-population*) over a set of types T. For our running example, suppose that $\mathcal{T}$ is uniform over $[50, 200]$, $n = 5$ and sampled types are $\vec{t} = (55, 80, 100, 120, 165)$, where lower types will provide better estimation in expectation (note that we use an arrow accent for n-length vectors).

(2) Workers each report their answers S, which depend on the ground truth $\boldsymbol{z}$, on their types, and on a random factor. $\boldsymbol{z}$ and S for our example are shown in Table 1 and Fig. 2.

Formally, a *noise model* is a function $\mathcal{Y} : Z \times T \to \Delta(Z)$. That is, the report of worker i is a random variable $\boldsymbol{s}_i$ sampled from $\mathcal{Y}(\boldsymbol{z}, t_i)$. We note that $\mathcal{T}$, $\mathcal{Y}$ and $\boldsymbol{z}$ together induce a distribution $\mathcal{Y}(\boldsymbol{z}, \mathcal{T})$ over answers (and thus over datasets), where $\boldsymbol{s} \sim \mathcal{Y}(\boldsymbol{z}, \mathcal{T})$ means we first sample a type $t \sim \mathcal{T}$ and then a report $\boldsymbol{s} \sim \mathcal{Y}(\boldsymbol{z}, t)$.

The data in our example (Table 1) was sampled from the noise model $\mathcal{Y}$ that is a multivariate independent Normal distribution with mean $\boldsymbol{z}$ and variance t_i. This is known as *Additive White Gaussian noise* (AWG, see (Diebold 1998)).

[1] It is ok if $\boldsymbol{s}_i$ is a partial vector, as long as there is enough intersection between pairs of workers.

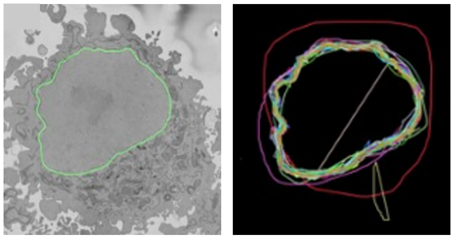

Figure 1: The Etch-a-Cell project. Left: an Electron Microscopy (EM) image of a cell. The real boundary of the tumor is marked in green. Right: multiple annotations by volunteers. Images taken from (Spiers et al. 2021).

$i \setminus j$		1	2	3	4	$d(\boldsymbol{s},\boldsymbol{z})$
	$t_i \setminus z_j$	80	0	40	-10	
1	55	81	6	41	-14	13.5
2	80	89	-6	35	4	84.5
3	100	105	-18	39	-5	243.7
4	120	68	9	62	-10	177.2
5	165	67	20	58	-20	248.2
$UA(S)$		82	2.2	47	-9	14.7

Table 1: An example of a dataset sampled from the AWG model. The bottom row is showing aggregated results using the unweighted mean.

Workers' competence Competent workers are close to the truth. More formally, given some ground truth $\boldsymbol{z}$ and a distance measure d, we define the *fault* (or *incompetence*) of a worker i as her expected distance from the ground truth, denoted $f_i(\boldsymbol{z}) := E_{\boldsymbol{s}_i \sim \mathcal{Y}(\boldsymbol{z},t_i)}[d(\boldsymbol{s}_i,\boldsymbol{z})]$.

g We denote by $\mu_{\mathcal{T}}(\boldsymbol{z}) := E_{\boldsymbol{s} \sim \mathcal{Y}(\boldsymbol{z},\mathcal{T})}[d(\boldsymbol{s},\boldsymbol{z})]$ the mean fault, omitting $\mathcal{T}$ and/or $\boldsymbol{z}$ when clear from the context.

Distance measures can often be derived from an inner product. Formally, consider an arbitrary symmetric inner product space $(Z, \langle \cdot,\cdot \rangle)$. This induces a norm $\|\boldsymbol{x}\|^2 := \langle \boldsymbol{x},\boldsymbol{x} \rangle$ and a distance measure $d(\boldsymbol{x},\boldsymbol{y}) := \|\boldsymbol{x}-\boldsymbol{y}\|^2$ (not necessarily a metric). A special case of interest is the normalized Euclidean product on $Z = \mathbb{R}^m$, defined as $\langle \boldsymbol{x},\boldsymbol{y} \rangle_E := \frac{1}{m}\sum_{j \leq m} x_j y_j$; and the corresponding *normalized squared Euclidean distance* (NSED), a natural way to capture the dissimilarity of two items (Carter, Morris, and Blashfield 1989). Note that the fault of a worker in the AWG model under NSED is her *variance*, as $f_i(\boldsymbol{z}) = t_i$ for any $\boldsymbol{z}$.

Aggregation Given an instance $\langle S, \boldsymbol{z} \rangle$, an aggregation function returns predicted labels $\hat{\boldsymbol{z}}$. We define the *error* as $d(\boldsymbol{z},\hat{\boldsymbol{z}})$. For example *unweighted aggregation* in the real-valued domain simply returns the mean of workers' answers. The goal of truth discovery is to find algorithms that return labels with low expected error, see Table 1.

When the type of every worker is known, for many noise models there are accurate characterizations of the optimal aggregation functions. For example, the best linear unbiased estimator under the AWG model with NSED is taking the mean of workers' answers, inversely weighted by their variance (Aitkin 1935).

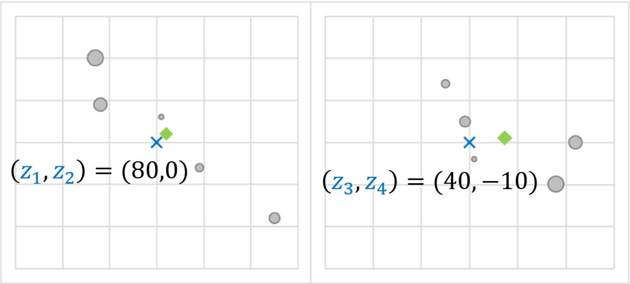

Figure 2: A graphical representation of Table 1 (for obvious reasons we use two 2-dimensional plots instead of a 4-dimensional one). The blue X marks the ground truth. Workers' reports are marked by gray circles, whose size is proportional to t_i (so smaller circles tend to be closer to the truth). The mean UA(S) is marked by a green diamond.

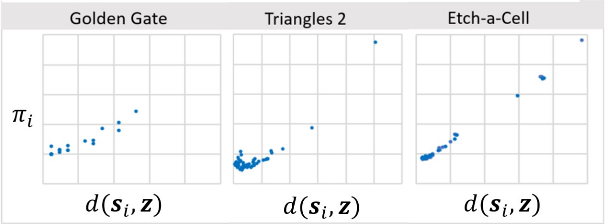

Figure 3: Realized average distance π_i vs. error. Each point is a worker.

Fault Estimation

Our key approach is relying on estimating f_i using the average distance of worker i from all other workers. Formally, we define $d_{ii'} := d(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$, and the *average pairwise distance* is

$$\pi_i := \frac{1}{n-1} \sum_{i' \in N \setminus \{i\}} d_{ii'}. \qquad (1)$$

Next, we analyze the relation between $\pi_i = \pi_i(S)$ (which is a random variable) and f_i, which is an inherent property that is deterministically induced by the worker's type. For an element $\boldsymbol{s} \in Z$ we consider the induced noise variable $\boldsymbol{\epsilon_s} := \boldsymbol{s} - \boldsymbol{z}$. We denote by $\tilde{\mathcal{Y}}(\boldsymbol{z},t)$ the distribution of $\boldsymbol{\epsilon_s}$ (where $\boldsymbol{s} \sim \mathcal{Y}(\boldsymbol{z},t)$). Thus under NSED we have that $d(\boldsymbol{s},\boldsymbol{z}) = \|\boldsymbol{\epsilon_s}\|^2$.

We define $\boldsymbol{b}_i(\boldsymbol{z}) := E_{\boldsymbol{\epsilon}_i \sim \tilde{\mathcal{Y}}(\boldsymbol{z},t_i)}[\boldsymbol{\epsilon}_i]$ as the *bias* of a type i worker, and $\boldsymbol{b}_{\mathcal{T}}(\boldsymbol{z}) := E_{\boldsymbol{\epsilon} \sim \tilde{\mathcal{Y}}(\boldsymbol{z},\mathcal{T})}[\boldsymbol{\epsilon}]$ as the mean bias of the proto-population. E.g. in Euclidean space $\boldsymbol{b}_i(\boldsymbol{z})$ is a vector where $b_{ij}(\boldsymbol{z}) > 0$ if i tends to overestimate the answer of question j, and negative values mean underestimation.

Our main conceptual result is an approximately linear connection between the expectations of π_i and $d(\boldsymbol{s}_i,\boldsymbol{z})$.

Theorem 1 (Anna Karenina Principle)**.**

$$E_{S \sim \mathcal{Y}(\boldsymbol{z},\mathcal{T})^n}[\pi_i | t_i, \boldsymbol{z}] = f_i(\boldsymbol{z}) + \mu_{\mathcal{T}}(\boldsymbol{z}) - 2\left\langle \boldsymbol{b}_i(\boldsymbol{z}), \boldsymbol{b}_{\mathcal{T}}(\boldsymbol{z}) \right\rangle.$$

We can also see this linear relation in three datasets (with different labels and distance measures) on Fig. 3.

The proof is rather straight-forward, and is relegated to the full version of the paper. In particular it shows by direct computation that the expectation of $d_{ii'}$ for every pair of workers is

$$E[d_{ii'}|t_i, t_{i'}, \boldsymbol{z}] = f_i(\boldsymbol{z}) + f_{i'}(\boldsymbol{z}) - 2\left\langle \boldsymbol{b}_i(\boldsymbol{z}), \boldsymbol{b}_{i'}(\boldsymbol{z})\right\rangle. \quad (2)$$

Despite (or perhaps because of) its simplicity, the principle above is highly useful for estimating workers' competence. If π_i is roughly linearly increasing in f_i, a naïve approach to estimate f_i from the data is by setting $\hat{f}_i$ to be some increasing function of π_i.

However there are several obstacles we need to overcome in order to get theoretical guarantees.

In particular: concentration bounds; estimation of μ; and the biases that appear in the last term; all of which will be tackled in the next sections.

For concreteness, we assume in the remainder of the paper, except where explicitly stated otherwise, that the inner product space $(Z, \langle\rangle)$ is $\mathbb{R}^m$, and d is NSED.

Concentration bounds How far is the empirical average π_i from its expectation? We show that when the noise on all questions is independent and bounded, the probability of a large estimation error decreases linearly with the sample size $\min\{n, m\}$.

Domain-independent Bounds

What can be said without symmetry or other assumptions on the model? We argue that we can at least tell particularly poor workers from good workers.

Corollary 2. *Consider a "bad" worker i^* with $f_{i^*} > 9\mu_{\mathcal{T}}$, and a "good" worker i^{**} with $f_{i^{**}} < \mu_{\mathcal{T}}$. Then $E[\pi_{i^*}] > E[\pi_{i^{**}}]$. No better separation is possible (i.e. there is an instance where all the inequalities become equalities).*

The proof relies on the following lemma, which is itself a corollary of the Anna Karenina principle (Thm. 1) and the Cauchy-Schwarz inequality:

Lemma 3. *For any worker i and any $\gamma \geq 0$, if $f_i = \gamma\mu$, then $E[\pi_i|t_i] \in (1 \pm \sqrt{\gamma})^2\mu$.*

Proof of Corollary 2. By Lemma. 3, $E[\pi_{i^{**}}] \leq 4\mu < E[\pi_{i^*}]$.

Without further assumptions, this condition is tight. To see why, consider a population on $\mathbb{R}$ where $z = 0$. The good worker i^{**} provides the fixed report $s_{i^{**}} = -1$, the poor worker i^* provides the fixed report $s_{i^*} = 3 - \delta$. However the measure of types $t_{i^*}, t_{i^{**}}$ in $\mathcal{T}$ is 0, and w.p. 1 type t' is selected with a fixed report $s = 1$. Note that $\mu = (s - z) = 1$, and thus $f_{i^{**}} = 1 = \mu$ whereas $f_{i^*} = 9 = 9\mu$.

However, the reports of i^*, i^{**} are completely symmetric around 1, in the absence of more workers there is no way to distinguish between these two workers, by their disparity or otherwise. □

In the special case of interest when there are only two types of workers (a situation known as "Hammer-spammer" (Karger, Oh, and Shah 2011)), Lemma 3 enables us to separate good from bad workers even more easily. This essentially depends on the fraction of bad workers and on their bias.

Symmetric Noise A trivial implication of Theorem is when the average worker is unbiased:

Corollary 4 (Anna Karenina principle for zero bias)**.** *If $\boldsymbol{b}_{\mathcal{T}} = 0$ then $E[\pi_i|t_i] = f_i + \mu_{\mathcal{T}}$ for all i.*

This means that given enough samples, we can retrieve workers' exact fault level with high accuracy, by setting $\hat{f}_i := \pi_i(S) - \hat{\mu}$. This will be important later on when we discuss aggregation.

What if we use other distance measures than NSED? Suppose that d is an *arbitrary distance metric* over space Z, $\boldsymbol{z} \in Z$ is the ground truth, and $\boldsymbol{s}_i \in Z$ is the report of worker i. f_i and π_i are defined as before. Intuitively, we say that the noise model $\mathcal{Y}$ is *symmetric* if for every point $\boldsymbol{x}$ there is an equally-likely point that is on "the other side" of $\boldsymbol{z}$ (note that this in particular implies zero bias).

Theorem 5 (Anna Karenina principle for symmetric noise and distance metrics)**.** *If d is any distance metric and $\mathcal{Y}$ is symmetric, then $\max\{\mu, f_i\} \leq E[\pi_i|t_i] \leq \mu + f_i$.*

An immediate corollary of Theorem 5 is that for poor workers with $f_i \geq \mu$, the average distance π_i is a 2-approximation for f_i (up to noise). See details and proof in the full version.

Domain-specific Results

Binary labels Kurvers *et al.* (2019) considered the average similarity of workers when answering a set of yes/no questions, and the type of a worker is her probability p_i to answer correctly independently over each question, a model known as the *one-coin* model or the *Dawid-Skene* model.

They showed that the (expected) average similarity is an increasing linear function of p_i.

Interestingly, the result from (Kurvers et al. 2019) can also be obtained directly from Theorem 1, by plugging in the Hamming distance (which is just NSED on the binary cube $\{-1, 1\}^d$ instead of $\mathbb{R}^d$). This result can also be easily extended to multiple-choice labels. For details see the full version.

Cosine similarity When label vectors are normalized, we have that $d(\boldsymbol{x}, \boldsymbol{y}) = 2(1 - cos(\boldsymbol{x}, \boldsymbol{y}))$, meaning that ranking workers by decreasing average cosine similarity (as suggested in (Kobayashi 2018)) is the same as ranking them by increasing average NSED. Our results above provide sufficient conditions for when this separates good workers from poor ones.

AWG Model and Maximum Likelihood

Since AWG has no bias, we know from Cor. 4 that $E[\pi_i(S)|t_i] = f_i + \mu$. Thus if we have a good estimate $\hat{\mu}$ of μ, setting $\hat{f}_i := \pi_i - \hat{\mu}$ is a reasonable heuristic. In this section we show that under a slight relaxation of the AWG model, tweaking the heuristic above provides the MLE for f_i.

Denote $\bar{\mu} := \frac{1}{2n}\sum_{i \in N}\pi_i(S)$, that is, half the average pairwise distance. Note that for unbiased workers, we have by Cor. 4 that $E[\frac{1}{n}\sum_{i \in N}\pi_i(S)] = 2\mu$, and thus $\bar{\mu}$ is an unbiased estimator of μ. We thus set $\hat{f}_i^{NP} := \pi_i - \bar{\mu}$, where NP stands for Naïve Proxy.

Computing the MLE From Eq. (2) and zero-bias, we have that $E[d_{ii'}|t_i, t_{i'}] = f_i + f_{i'}$. However even under the AWG model, the pairwise distances are correlated. For the analysis, we will neglect these correlations, and assume that the pairwise distances are all independent conditional on workers' types. More formally, under this *'pairwise' Additive White Gaussian* (pAWG) model, $d_{ii'} = f_i + f_{i'} + \epsilon_{ii'}$, where all of $\epsilon_{ii'}$ are sampled i.i.d. from a normal distribution with mean 0 and unknown variance. Ideally, we would like to find $\ddot{f} := (\hat{f}_i)_{i \in N}$ that minimize the estimation errors $(\epsilon_{ii'})$,[2] which is an Ordinary Least Squares (regression) problem.

We next show how to derive a closed form solution for the maximum likelihood estimator of the fault $\vec{f}$, also allowing for a regularization term with coefficient λ. The theorem might have independent interest as it allows us to estimate a matrix created from adding a vector to itself (an "outer sum" matrix) from its off-diagonal entries.

Theorem 6. *Let $\lambda \geq 0$, and $D = (d_{ii'})_{i,i' \in N}$ be an arbitrary symmetric nonnegative matrix. Then*

$$\textit{argmin}_{\ddot{f}} \sum_{i,i' \in N : i \neq i'} (\hat{f}_i + \hat{f}_{i'} - d_{ii'})^2 + \lambda \|\ddot{f}\|_2^2 = \frac{2(n-1)\vec{\pi} - \frac{8n(n-1)}{4n+\lambda-4}\bar{\mu}}{2n + \lambda - 4}.$$

Our main technical result in this paper follows as a direct corollary of the above theorem, when the matrix D represents pairwise distances:

Theorem 7. *For $\lambda = 4$, the regularized maximum likelihood estimator of $\vec{f}$ in the pAWG model is proportional to $\ddot{f}^{NP}$.*

That is, our heuristic estimate $\ddot{f}^{NP}$ is in fact the optimal solution of a regularized pAWG model.

By setting $\lambda = 0$ (no regularization) we get a slight variation of this heuristic $\ddot{f}^{ML} := \vec{\pi} - \frac{n}{n-1}\bar{\mu}$ (for 'maximum likelihood').

In Fig. 4 we compare implementations of our truth discovery algorithm (defined in Section , derived with different values of λ. As expected, more regularization leads to better performance on a small dataset, whereas the unregularized version is optimal in the limit.

Proof of Theorem 6 for $\lambda = 0$. Let P be a list of all $n^2 - n$ ordered pairs of $[n]$ (without the main diagonal) in arbitrary order. Setting $\lambda = 0$, we are left with the following least squares equation:

$$\min_{\ddot{f}} \sum_{i,i'} (\hat{f}_i + \hat{f}_{i'} - d_{ii'})^2 = \min_{\ddot{f}} \|A\ddot{f} - \boldsymbol{d}\|_2^2,$$

where A is a $|P| \times n$ matrix with $a_{ki} = 1$ iff $i \in P_k$, and $\boldsymbol{d}$ is a $|P|$-length vector with $d_k = d_{ii'}$ for $P_k = (i, i')$.

Fortunately, A has a very specific structure that allows us to obtain the above closed-form solution. Note that every row of A has exactly two '1' entries, in the row index i and column index i' of P_k; the total number of '1' is $2(n^2 - n)$; there are $2n-2$ ones in every column; and every two distinct columns i, i' share exactly two non-zero entries (at rows k s.t. $P_k = (i, i')$ and $P_k = (i', i)$). This means that $(A^T A)$ has $2n - 2$ on the diagonal and 2 in any other entry.

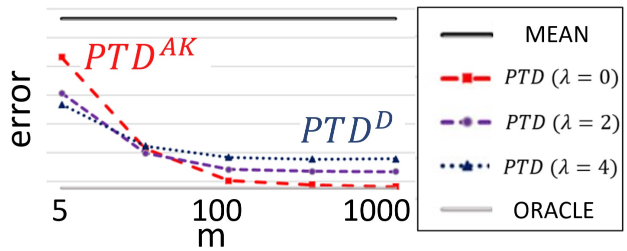

Figure 4: A comparison of the P-TD algorithm variants on a synthetic real-valued dataset. The x-axis shows the number of questions m, whereas the number of workers is fixed ($n = 5$). 'ORACLE' weighs workers according to their true competence.

The optimal solution for ordinary least squares is obtained at $\ddot{f}$ such that $(A^T A)\ddot{f} = A^T \boldsymbol{d}$. By the structure of A:

$$[A^T \boldsymbol{d}]_i = 2 \sum_{i' \neq i} d_{ii'} = 2\pi_i, \tag{3}$$

and
$$[(A^T A)\ddot{f}]_i = 2(n-2)\hat{f}_i + 2 \sum_{i' \in N} \hat{f}_{i'}. \tag{4}$$

Denote
$$\alpha := \sum_{i \in N} [A^T \boldsymbol{d}]_i = \sum_{i \in N} 2 \sum_{i' \neq i} d_{ii'} = 4n(n-1)\bar{\mu}, \tag{5}$$

then by Eq. (4):

$$\alpha = \sum_{i \in N} [(A^T A)\ddot{f}]_i = \sum_{i \in N} [2(n-2)\hat{f}_i + 2 \sum_{i' \in N} \hat{f}_{i'}] \tag{6}$$
$$= 2(n-2) \sum_{i \in N} f_i + 2 \sum_{i' \in N} f_{i'} (\sum_{i \in N} 1) = 4(n-1) \sum_{i \in N} \hat{f}_i.$$

We can now write the n linear equations as

$$2(n-1)\pi_i = [A^T \boldsymbol{d}]_i = [(A^T A)\ddot{f}]_i \qquad \text{(By Eqs. (3),(4))}$$
$$= (2n-4)\hat{f}_i + 2 \sum_{i' \in N} \hat{f}_{i'} \qquad \iff$$
$$(n-1)\pi_i = (n-2)\hat{f}_i + \sum_{i' \in N} \hat{f}_{i'} \qquad \iff$$
$$\hat{f}_i = \frac{\pi_i}{n-2} - \frac{1}{n-2} \sum_{i' \in N} \hat{f}_{i'} = \frac{\pi_i}{n-2} - \frac{1}{n-2} \frac{\alpha}{4(n-1)}$$
$$\overset{Eq. (5)}{=} \frac{n-1}{n-2}\pi_i - \frac{n}{n-2}\bar{\mu} = \frac{n-1}{n-2}\hat{f}_i^{ML},$$

as required, since $\frac{n-1}{n-2}$ is a constant. □

Aggregation

Our Proximity-based Truth Discovery (P-TD) algorithm is a direct adaptation of the Anna Karenina principle. The idea is very simple:

[2] We use $\ddot{f}$ instead of hat+arrow accent.

ALGORITHM 1: (P-TD^D) FOR REAL-VALUED DATA

```
Input: Dataset $S \in \mathbb{R}^{n \times m}$.
Output: Est. fault levels $\ddot{f} \in \mathbb{R}^n$; answers $\hat{\boldsymbol{z}} \in \mathbb{R}^m$.
Compute $d_{ii'} \leftarrow d(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$ for every pair of workers;
for each worker $i \in N$ do
    set $\pi_i \leftarrow \frac{1}{n-1} \sum_{i' \neq i} d_{ii'}$;            // Step 1
end
Set $\bar{\mu} \leftarrow \frac{1}{2n} \sum_{i \in N} \pi_i$;
for each worker $i \in N$ do
    Set $\hat{f}_i \leftarrow \pi_i - \frac{n}{n-1} \bar{\mu}$;            // Step 2
    Set $w_i \leftarrow \frac{1}{\hat{f}_i}$;
end
Set $\hat{\boldsymbol{z}} \leftarrow \frac{\sum_i w_i \boldsymbol{s}_i}{\sum_i w_i}$;            // Step 3
return $(\ddot{f}, \hat{\boldsymbol{z}})$;
```

1. Compute the average distance [or similarity] π_i of every worker;
2. Estimate fault [or competence] $\ddot{f}$ from $\vec{\pi}$;
3. Aggregate answers, giving higher weight to workers with low fault [high competence].

Our default implementation (denoted P-TD^D) simply sets weights proportional to the estimated competence, which is in turn proportional to the average similarity, as in (Kobayashi 2018; Kurvers et al. 2019).

As we make more assumptions on the structure of labels and the statistical model, we can use an appropriate Anna-Karenina theorem to improve Step 2, resulting in a domain-specific implementation P-TD^{AK}. For example, Alg. 1 shows the implementation for the real-valued domain, where Step 2 is based on $\ddot{f}^{ML}$ defined in Sec. , and Step 3 is based on (Aitkin 1935).

Lastly, we can iteratively repeat the process by computing the *weighted* average distance to other workers. This iterative P-TD algorithm is denoted by IP-TD.

Empirical Evaluation

Algorithms We compare the predicted label accuracy of our algorithms (P-TD^D,P-TD^{AK},IP-TD) to unweighted aggregation (UA); to three general-domain algorithms: MAS (Braylan and Lease 2020), TOP2 and EXP (Kawase, Kuroki, and Miyauchi 2019); and to domain-specific algorithms: CRH (Li et al. 2014b), IBP (Karger, Oh, and Shah 2011), DS (Dawid and Skene 1979), EVD (Parisi et al. 2014), CATD (Li et al. 2014a), GTM (Zhao and Han 2012), and KDE (Wan et al. 2016).

Datasets We used the following datasets from five different domains. We write the used distance measure in each domain in brackets.

Categorical (Hamming distance): GG, DOGS, FLAGS (Shah and Zhou 2015); Predict (Mandal, Radanovic, and Parkes 2020) (we used data from Oct.8, under all four treatments); and all six categorical datasets from (Kawase, Kuroki, and Miyauchi 2019). We used weighted majority for aggregation.

Real-valued (NSED): BUILDINGS (collected for this paper); TRI (Hart et al. 2018); and EMO (Snow et al. 2008). Answers aggregated using weighted mean.

Ranking (Kendall-tau): DOTS and PUZZ contain subjective rankings of four images of dots / 8-puzzle boards, according to the number of dots they contain / number of steps from solution (Mao, Procaccia, and Chen 2013). We also extracted the ranking information from BUILDINGS. For aggregation, we used nine different ordinal voting rules, see full version for details.

Language (GLEU): The TRANSL dataset contains English translations of Japanese sentences (Braylan and Lease 2020). The distance measure we used is GLEU, and there is no aggregation (best worker is selected).

Outlines (Jaccard): The Etch-a-Cell dataset contains bitmaps of the outline of a tumor in 2D slices of a cell (Spiers et al. 2021) (see Fig. 1). We use Jaccard distance on the filled shapes, and aggregate labels using pixelwise-majority.

In addition we generated synthetic datasets using the AWG model (real-valued); the one-coin model (categorical); and Mallows model (ranking). In the HS datasets there are 20% 'hammers'.

A detailed description of algorithms' implementation and datasets is in the full version of the paper.

To obtain robust results we sampled n workers and m questions without repetition from each dataset (real or synthetic), and repeated the process at least 1000 times for every combination.

Evaluation The *error* of every algorithm is the distance (as specified above) to the ground truth, averaged over all samples of certain size of a particular dataset.

In the tables, we compute for each algorithm its *Relative Improvement* $RI(Alg) := \frac{Err(Alg) - Err(\text{UA})}{Err(Alg) + Err(\text{UA})}$, where UA serves as a baseline. Thus RI is in the range $[-1, 1]$ where negative numbers mean improvement over UA.

In some cases we see that one algorithm has slightly higher average error (on the graphs) but lower RI, or that the gap in RI is more substantial. This is since the graphs average over instances of varying difficulty, so instances with high baseline error have more effect.

Results Fig. 5 and Table 2 (and more in the full version) show results on categorical and real-valued data, where there are many specialized algorithms. We can see that there is no single 'state-of-the-art', as algorithms that do well on some datasets may have poor performance on other data, or for a different number of voters/questions. This is especially true for the three general-domain algorithms.

Yet for moderate n and m, all three versions of our P-TD consistently provide good results over almost all datasets, usually beating the three general-domain algorithms, and doing roughly at par with the best specialized ones. We can

Categorical	k	CRH	IBP	DS	EVD	EXP	TOP2	MAS	P-TD^D	P-TD^{AK}	iP-TD
SYN.HS	2	-14%	+6%	**-23%**	-20%	-15%	-13%	-17%	-21%	**-24%**	**-24%**
SYN.N	2	-4%	+12%	**-9%**	-6%	-7%	**-8%**	-5%	**-7%**	**-9%**	**-9%**
GG	2	-8%	-15%	**-19%**	-10%	-14%	**-20%**	-14%	-14%	**-18%**	-18%
Pred.T1	2	**-1%**	+9%	**-1%**	**-1%**	**-1%**	+0%	**-1%**	**-1%**	**-1%**	**-1%**
Pred.T2	2	**-0%**	+10%	+0%	**-1%**	**-1%**	+1%	+0%	**-0%**	**-1%**	**-1%**
Pred.T3	2	**-1%**	+8%	**-1%**	**-1%**	-0%	+1%	**-1%**	**-2%**	**-2%**	**-2%**
Pred.T4	2	-1%	+7%	-2%	**-4%**	-1%	-1%	-2%	**-3%**	**-3%**	**-3%**
SYN	4	-18%	–	–	–	-22%	-29%	-28%	-33%	-33%	**-41%**
DOGS	10	-1%	–	–	–	-2%	+0%	**-3%**	**-4%**	**-4%**	**-4%**
FLAGS	4	-17%	–	–	–	-29%	**-31%**	-27%	-21%	-22%	**-30%**
Chinese	2	-2%	–	–	–	-1%	**-3%**	**-4%**	**-4%**	**-4%**	**-5%**
English	2	**-1%**	–	–	–	**-1%**	**-1%**	**-1%**	**-1%**	**-1%**	**-1%**
IT	4	-3%	–	–	–	-4%	**-5%**	**-6%**	**-5%**	**-5%**	**-7%**
Medicine	4	-3%	–	–	–	-13%	**-16%**	-9%	-8%	-8%	-12%
Pokemon	6	-11%	–	–	–	-25%	**-27%**	-17%	-22%	-22%	**-27%**
Science	5	**-1%**	–	–	–	+1%	+1%	**-1%**	**-3%**	**-3%**	**-3%**
Real-valued		CRH	CATD	GTM	KDE	EXP	TOP2	MAS	P-TD^D	P-TD^{AK}	iP-TD
SYN.N	–	-23%	-27%	-14%	+6%	-7%	-11%	+4%	-20%	**-41%**	-25%
BUILD	–	-9%	+5%	-9%	+1%	-8%	+12%	+4%	-7%	-7%	**-11%**
TRI1	–	**-19%**	-7%	-14%	-4%	-17%	-3%	-8%	-17%	-10%	**-20%**
TRI2	–	-9%	-6%	-5%	-4%	**-11%**	**-12%**	+1%	-7%	-4%	**-11%**
EMO	–	+1%	+18%	+1%	+26%	+5%	+22%	+3%	+0%	+5%	+2%

Table 2: Results (RI) on categorical and real-valued datasets, with $n = 10$ workers and $m = 15$ questions . The best result in each row is underlined, and results that are not statistically different (within 95% confidence interval in a paired t-test) are marked in bold. Results in gray are worse than unweighted aggregation.

also see that on synthetic real-valued data with Gaussian noise, the provably-optimal P-TD^D is also best in practice.

On real datasets, our iP-TD is usually better, and as n and m increase sometimes one of the specialized algorithms takes over. Intuitively, real datasets may often have a some correlation in poor workers' errors. Iterative algorithms, such as our iP-TD and some of the existing algorithms, are able to overcome this since they gradually rely more on the largest and most consistent set of workers.

The real strength of our approach shows when labels are more complex. Table 3 and Fig. 7 show how our simple algorithms are consistently better than the other three algorithms both on ranking data and on both complex annotation tasks. Results also show that P-TD yields substantial improvement regardless of the voting rule in use. Moreover, while the other algorithms work better on some datasets, they are highly unstable and often perform worse than the baseline.

Conclusion

Average proximity can be used as a general scheme to estimate workers' competence in a broad range of truth discovery and crowdsourcing scenarios. Due to the "Anna Karenina principle," we expect the answers of competent workers to be much closer to others, than those of incompetent workers, even under very weak assumptions on the domain and the noise model. Under more explicit assumptions, the average distance accurately estimates the true competence.

The above results suggest an extremely simple, general and practical algorithm for truth-discovery (the P-TD algorithm), that weighs workers by their average proximity to others, and can be combined with most aggregation methods. This is particularly useful in the context of existing crowdsourcing systems where the aggregation rule may be subject to constraints due to legacy, simplicity, explainability, legal, or other considerations (e.g. a voting rule with certain axiomatic properties). In addition, average proximity is simple and flexible enough so we can modify it to deal with challenges outside the scope of the current paper, such as partial data (Dalvi et al. 2013; Karger, Oh, and Shah 2011; Li et al. 2014a); [3] semi-supervised learning (Yin and Tan 2011); or worker's competence that varies across task types (Braylan and Lease 2020).

Despite its simplicity, the P-TD algorithm substantially improves the outcome compared to unweighted aggregation. It is also competitive with other, more sophisticated algorithms, especially in the common case of moderate input size. We thus conclude that the average similarity heuristic is indeed a frustratingly easy—and practical—tool for crowdsourcing.

An obvious shortcoming of P-TD is that a group of workers that submit similar labels (e.g. by acting strategically) can boost their own weights. Future work will consider how to identify and/or mitigate the affect of such groups.

[3] Preliminary experiments with partial data show similar results if workers still have nontrivial intersection with some other workers. This is so that average similarity can be reasonably estimated.

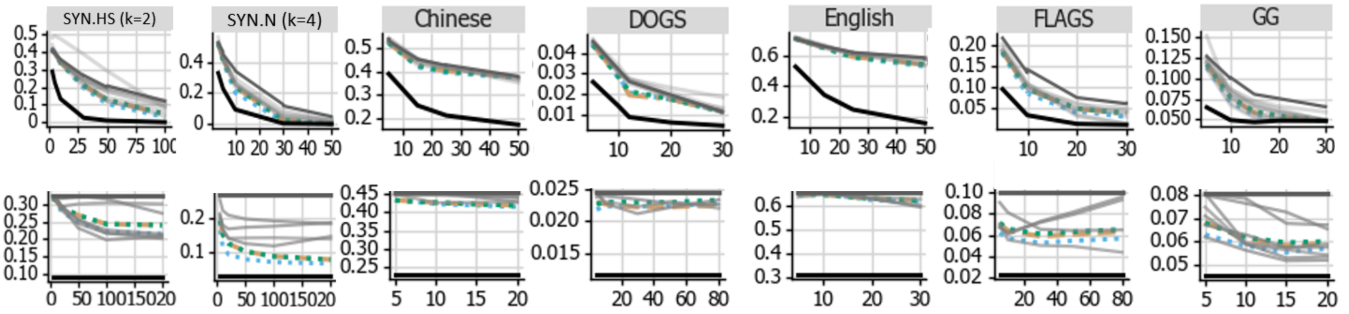

Figure 5: The error on some of the real world datasets as we increase the number of voters n with fixed $m = 15$ as in the table (top row), or vary the number of questions m with fixed $n = 10$ (bottom row). The black line is the error of an *Oracle* who knows the true fault values f and uses optimal weights. The thin gray lines are all competing algorithms.

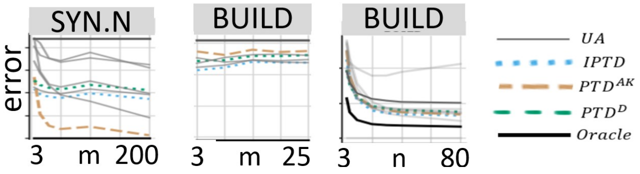

Figure 6: The error on real-valued datasets as we vary m (keeping $n = 10$) or n (keeping $m = 15$).

Ranking	v. rule	TOP2	MAS	P-TD	IP-TD
SYN.HS	Borda	**-24%**	-14%	-4%	-10%
SYN.N	Borda	-3%	+1%	**-4%**	**-6%**
BUILD	Borda	+0%	+0%	**-0%**	+0%
DOTS3	Borda	+7%	+2%	**-1%**	+1%
DOTS5	Borda	+9%	+4%	**-1%**	**-1%**
DOTS7	Borda	+13%	+5%	**-3%**	**-2%**
DOTS9	Borda	+7%	-1%	-6%	**-10%**
PUZZ5	Borda	+1%	+8%	**-2%**	**-2%**
PUZZ7	Borda	-7%	-18%	-15%	**-20%**
PUZZ9	Borda	+48%	+14%	**-5%**	-1%
PUZZ11	Borda	+6%	+3%	**-3%**	**-3%**
SYN.N	Plu.	-2%	**-5%**	**-4%**	**-5%**
BUILD	Plu.	+20%	-2%	**-6%**	**-6%**
DOTS3	Plu.	+14%	+2%	**-2%**	**-1%**
PUZZ5	Plu.	+12%	+1%	**-3%**	**-2%**
SYN.N	Cop.	**-37%**	-25%	-27%	-29%
BUILD	Cop.	-7%	-17%	**-27%**	**-27%**
DOTS3	Cop.	+0%	**-1%**	**-0%**	**-0%**
PUZZ5	Cop.	-12%	-16%	**-19%**	**-18%**
Complex					
TRAN	best v.	-3%	-3%	**-4%**	**-4%**
ETCH	best v.	-39%	-39%	**-43%**	**-42%**
ETCH	bit. mj.	-2%	**-4%**	-2%	-2%

Table 3: Results (RI) for rankings datasets, under three different voting rules ($n = 10$, four ranked alternatives), and on the other complex annotation datasets. EXP removed due to space constraints, and since it did not have the best results in any line.

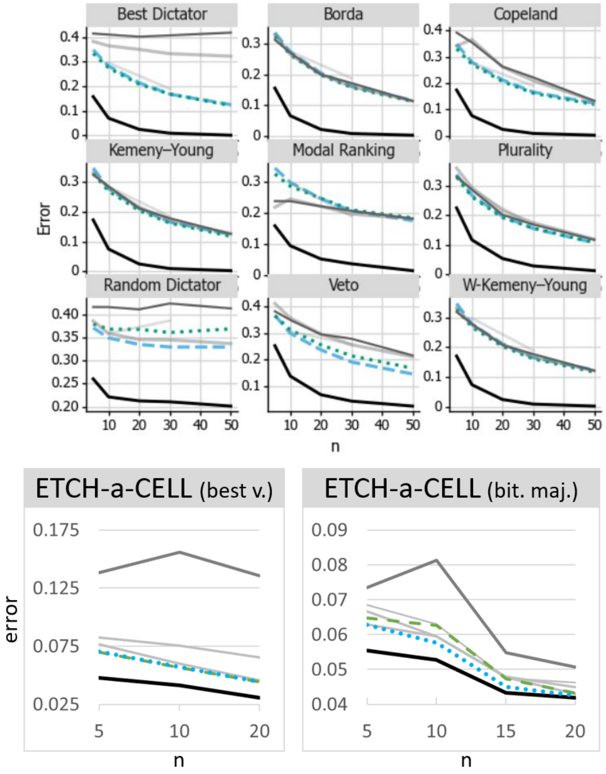

Figure 7: Top: error on DOTS3 under nine different voting rules. Bottom: error on Etch-a-Cell as number of workers grows.

Acknowledgements

This research was supported by the Israel Science Foundation (ISF; Grant No. 2539/20).

Previous versions of this paper have been rejected from 7 (seven!) AI conferences. The current version is much improved thanks to the comments, suggestions, and references provided by the (many) reviewers along the way.

References

Aitkin, A. 1935. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55: 42–48.

Aydin, B. I.; Yilmaz, Y. S.; Li, Y.; Li, Q.; Gao, J.; and Demirbas, M. 2014. Crowdsourcing for multiple-choice question answering. In *Proceedings of the 26th IAAI Conference*.

Braylan, A.; and Lease, M. 2020. Modeling and Aggregation of Complex Annotations via Annotation Distances. In *Proceedings of The Web Conference 2020*, 1807–1818.

Caragiannis, I.; Procaccia, A. D.; and Shah, N. 2013. When do noisy votes reveal the truth? In *Proceedings of the 14' Conference on Economics and Computation (EC'13)*, 143–160. ACM.

Carter, R. L.; Morris, R.; and Blashfield, R. K. 1989. On the partitioning of squared Euclidean distance and its applications in cluster analysis. *Psychometrika*, 54(1): 9–23.

Condorcet, M. J. 1785. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Trans. lain McLean and Fiona Hewitt. Paris.

Dalvi, N.; Dasgupta, A.; Kumar, R.; and Rastogi, V. 2013. Aggregating crowdsourced binary ratings. In *Proceedings of the Web Conference (WWW'13)*, 285–294.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

Deng, J.; Russakovsky, O.; Krause, J.; Bernstein, M. S.; Berg, A.; and Fei-Fei, L. 2014. Scalable multi-label annotation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'14)*, 3099–3102. ACM.

Diebold, F. X. 1998. *Elements of forecasting*. South-Western College Pub.

Gadiraju, U.; Fetahu, B.; Kawase, R.; Siehndel, P.; and Dietze, S. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction*, 24(4): 1–26.

Gao, C.; and Zhou, D. 2013. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*.

Grofman, B.; Owen, G.; and Feld, S. L. 1983. Thirteen theorems in search of the truth. *Theory and Decision*, 15(3): 261–278.

Hart, Y.; Dillon, M. R.; Marantan, A.; Cardenas, A. L.; Spelke, E.; and Mahadevan, L. 2018. The statistical shape of geometric reasoning. *Scientific reports*, 8(1): 12906.

Huang, S.-W.; and Fu, W.-T. 2013. Enhancing reliability using peer consistency evaluation in human computation. In *Proceecings of the ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW'13)*, 639–648.

Karger, D. R.; Oh, S.; and Shah, D. 2011. Iterative learning for reliable crowdsourcing systems. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'11)*, 1953–1961.

Kawase, Y.; Kuroki, Y.; and Miyauchi, A. 2019. Graph mining meets crowdsourcing: Extracting experts for answer aggregation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'19)*.

Kobayashi, H. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proccedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*, 4165–4176.

Kurvers, R. H.; Herzog, S. M.; Hertwig, R.; Krause, J.; Moussaid, M.; Argenziano, G.; Zalaudek, I.; Carney, P. A.; and Wolf, M. 2019. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science advances*, 5(11): eaaw9011.

Li, J.; Baba, Y.; and Kashima, H. 2018. Incorporating worker similarity for label aggregation in crowdsourcing. In *International Conference on Artificial Neural Networks*, 596–606. Springer.

Li, Q.; Li, Y.; Gao, J.; Su, L.; Zhao, B.; Demirbas, M.; Fan, W.; and Han, J. 2014a. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4): 425–436.

Li, Q.; Li, Y.; Gao, J.; Zhao, B.; Fan, W.; and Han, J. 2014b. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD'14*.

Li, X.; Dong, X. L.; Lyons, K.; Meng, W.; and Srivastava, D. 2012. Truth finding on the deep web: is the problem solved? *Proceedings of the VLDB Endowment*, 6(2): 97–108.

Li, Y.; Gao, J.; Meng, C.; Li, Q.; Su, L.; Zhao, B.; Fan, W.; and Han, J. 2016. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2): 1–16.

Mandal, D.; Radanovic, G.; and Parkes, D. C. 2020. The Effectiveness of Peer Prediction in Long-Term Forecasting. In *Proceedings of the Conference on Artificial Intelligence (AAAI'20)*.

Mao, A.; Procaccia, A. D.; and Chen, Y. 2013. Better human computation through principled voting. In *Proceedings of the Conference on Artificial Intelligence (AAAI'13)*.

McLaughlin, E. 2014. Image Overload: Help us sort it all out, NASA requests. CNN.com. Retrieved at 18/9/2014.

Parisi, F.; Strino, F.; Nadler, B.; and Kluger, Y. 2014. Ranking and combining multiple predictors without labeled data. *PNAS*, 111(4): 1253–1258.

Prelec, D.; Seung, H. S.; and McCoy, J. 2017. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638): 532.

Shah, N. B.; and Zhou, D. 2015. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'15)*, 1–9.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, 254–263.

Spiers, H.; Songhurst, H.; Nightingale, L.; De Folter, J.; Community, Z. V.; Hutchings, R.; Peddie, C. J.; Weston, A.; Strange, A.; Hindmarsh, S.; et al. 2021. Deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations. *Traffic*, 22(7): 240–253.

Von Ahn, L.; and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8): 58–67.

Vuurens, J.; de Vries, A. P.; and Eickhoff, C. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *ACM SIGIR Workshop on CIR' 11*, 21–26.

Wais, P.; Lingamneni, S.; Cook, D.; Fennell, J.; Goldenberg, B.; Lubarov, D.; Marin, D.; and Simons, H. 2010. Towards building a high-quality workforce with mechanical turk. *NeurIPS workshop*, 1–5.

Wan, M.; Chen, X.; Kaplan, L.; Han, J.; Gao, J.; and Zhao, B. 2016. From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (SIGKDD'16)*, 1885–1894.

Witkowski, J.; Bachrach, Y.; Key, P.; and Parkes, D. 2013. Dwelling on the negative: Incentivizing effort in peer prediction.

In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1.

Xiao, H.; Gao, J.; Wang, Z.; Wang, S.; Su, L.; and Liu, H. 2016. A truth discovery approach with theoretical guarantee. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (SIGKDD'16)*, 1925–1934.

Yin, X.; and Tan, W. 2011. Semi-supervised truth discovery. In *Proceedings of the Web Conference (WWW'11)*, 217–226.

Zhang, Y.; Chen, X.; Zhou, D.; and Jordan, M. I. 2016. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1): 3537–3580.

Zhao, B.; and Han, J. 2012. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 1817.