Frustratingly Easy Truth Discovery for Rank Aggregation[1]

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
{reshefm, oamir, omerbp}@ie.technion.ac.il, {tsviel,galcohensius}@gmail.com
[2] RPI, 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. If we further assume conditional independence of workers, there is a linear or approximately linear relation between workers' quality and their average similarity. This applies in particular when labels are ordinal rankings.

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.

We provide a result that may be of independent interest, showing that the optimal way to aggregate rankings under Condorcet noise model with known heterogeneous competence is a weighted Kemeny voting rule with specific weights.

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

1 Introduction

Consider a standard crowdsourcing task such as identifying which images contain a person or a car [10], or identifying the location in which pictures were taken [26]. 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 [33, 34], 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* [7], 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 [31].

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 [18] suggests a 'frustratingly easy' algorithm that ranks workers by their *average cosine similarity* to

others in a text summarization task; and Kurvers et al. [19] 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* [32, 15] to identify outliers, and in peer prediction to incentivize effort [35].

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 in Section 2, we prove a formal version of the Anna Karenina principle and show how it can be used to identify poor workers in Section 3 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 4 we explain how to leverage the Anna Karenina principle for aggregation using a simple algorithm (P-TD). In addition, we prove that if the competence of all workers is known, then the optimality of the Kemeny-Young voting rule under the Condorcet noise model can be extended to heterogeneous voters (with appropriate weights).

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: https://tinyurl.com/2p923tzv.

The full version of the AAAI23 paper is available on arXiv: https://arxiv.org/abs/1905.00629.

Related work

The Condorcet Jury Theorem [7] was perhaps the first formal treatment of truth discovery, and extensions to experts with heterogeneous competence levels were surveyed by Grofman et al. [14]. 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 [23]). We should note that *self-reporting* of accuracy often leads to poor results [12, 29].

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 [18] proved that cosine similarity approximates a known kernel density estimator. Kurvers et al. [19] 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 [9], with many follow-ups [16, 13, 2, 37, 41, 22].

Another class of algorithms uses spectral methods to infer the competence and/or other latent variables from the covariance matrix of the workers [28, 40], or from their pairwise Hamming similarity [20]. 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 [25, 5].

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 [4] 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 [17], 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.

Rank aggregation The idea that voting rules can be used as a statistical tool to extract the most likely truth goes back at least to Condorcet: He suggested the model where every voter/expert swaps each pair of alternatives with a fixed independent probability [7]. The optimal solution to this particular problem is given by the Kemeny voting rule, as shown by Young [39]. Followup work considered other noise models [11, 36].

More recently, Procaccia et al. considered voting rules as a possible solution to *adversarial noise* [30].

In all these works, voters are assumed to be equally competent a-priori. However almost every voting rule has a natural weighted variation (e.g. by duplicating voters with high competence), so a good method to estimate competence is likely to boost the performance of any such voting rule.

2 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*.[2]

Distance measures We assume there is a distance measure (not necessarily a metric) $d : Z \times Z \to \mathbb{R}_+$ over pairs of labels.

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

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 [6].

Rankings Let $\mathcal{L}(C)$ be the set of all permutations over a set of candidates C. Any ranking (permutation) $L \in \mathcal{L}(C)$ corresponds to a binary vector $\boldsymbol{s} \in \{0,1\}^m$ where $m = \binom{|C|}{2}$ (i.e., each dimension is a pair of candidates), which we denote by $\boldsymbol{s}_L \in \{0,1\}^m$. If there is $L \in \mathcal{L}(C)$ s.t. $\boldsymbol{s}_L = \boldsymbol{s}$ then we denote it $L_{\boldsymbol{s}}$. In particular we denote by $L_{\boldsymbol{z}}$ the ground truth order over C.

A natural metric over rankings is the [normalized] Kendall-tau distance (a.k.a swap distance), that coincides with the [normalized] Hamming distance on the binary representation. That is, $d_{KT}(L, L') = d_{Hamming}(\boldsymbol{s}_L, \boldsymbol{s}_{L'})$. Hamming distance, in turn, is a special case of NSED.

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 Z is the set of permutations over 5 alternatives $\mathcal{L}(\{a, b, c, d, e\})$, $\mathcal{T}$ is uniform over $[0.3, 0.8]$, $n = 5$ and sampled types are $\vec{t} = (0.35, 0.4, 0.4, 0.5, 0.75)$, where lower is more accurate.

(2) Workers each report their answers S, which depend on the ground truth $\boldsymbol{z}$ (the identity permutation in our example), on their types, and on a random factor. We next describe an example where labels are rankings. See a different example with real-valued answers and Gaussian noise in full version.

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 Mallows distribution centered on a ranking $L_{\boldsymbol{z}}$ with individual parameter $\phi_i := t_i$. This means worker i reports every order $L_i = L_{\boldsymbol{s}_i} \in \mathcal{L}$ with probability proportional to $\phi_i^{-d(L_{\boldsymbol{z}}, L_i)}$, where $d(\cdot, \cdot)$ is the Kendall-tau distance. Thus for low ϕ_i the worker is very likely to return the true permutation with few errors, and for ϕ_i close to 1, the distribution is almost uniform.

Workers' competence Competent workers are close to the truth, *in expectation*. 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 $f_i(\boldsymbol{z}) := E_{\boldsymbol{s}_i \sim \mathcal{Y}(\boldsymbol{z}, t_i)}[d(\boldsymbol{s}_i, \boldsymbol{z})]$.

In the case of Mallows distribution with parameter ϕ_i, it is known that it is equivalent to Condorcet Noise model, which swaps every pair of items with probability p_i, where $\phi_i = \frac{p_i}{1 - p_i}$, see [24].

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

Aggregation Given an instance $\langle S, \boldsymbol{z} \rangle$, an aggregation function returns predicted labels $\hat{\boldsymbol{z}}$. We define the *error* as $d(\boldsymbol{z}, \hat{\boldsymbol{z}})$. *Unweighted aggregation* is usually trivial in binary domain (Majority) and in real valued domains (mean or median). However in the ranking domain there are countless

	$t_i = \phi_i$	p_i	L_i	$d(L_i, L_{\boldsymbol{z}})$
v_1	0.35		$\underline{b}acd\underline{e}$	1
v_2	0.4		$abc\underline{e}\underline{d}$	1
v_3	0.4		$ab\underline{d}c\underline{e}$	1
v_4	0.5		$ad\underline{b}c\underline{e}$	2
v_5	0.75		$\underline{ba}\underline{e}\underline{d}\underline{c}$	4
$Borda$			$ab\underline{d}ce$	1

$\Rightarrow$

d_{KT}	v_1	v_2	v_3	v_4	v_5
v_1	–	2	2	3	3
v_2	2	–	2	3	3
v_3	2	2	–	1	3
v_4	3	3	1	–	4
v_5	3	3	3	4	–
π_i	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{4}{12}$	$\frac{11}{24}$	$\frac{13}{24}$

Table 1: Left: an example of a ranking dataset sampled from a Mallows distribution, with ground truth $L_{\boldsymbol{z}} = abcde$. The bottom row is showing aggregated results using the (unweighted) Borda rule. Misplaced items in each ranking are underlined. The rightmost column shows the *error*, i.e. the Kendall-tau distance of every voter from the ground truth. Right: the pairwise KT distance matrix for the same five workers, and the average distance π.

voting rules, and each one of them can be used as a valid aggregation rule. Most common rules also have natural weighted extensions.

We do not aim to find voting rules that are better than others, but to estimate workers' competence and weigh them accordingly, so that the result under *any* voting rule would improve.

3 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

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

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 $\boldsymbol{\epsilon_s} := \boldsymbol{s} - \boldsymbol{z}$.[3] 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 see this linear relation in both synthetic and real datasets (with different labels and distance measures) on Fig. 1.

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

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

[3] When labels are rankings, $\boldsymbol{s}$, $\boldsymbol{z}$ etc. are their representations as binary vectors.

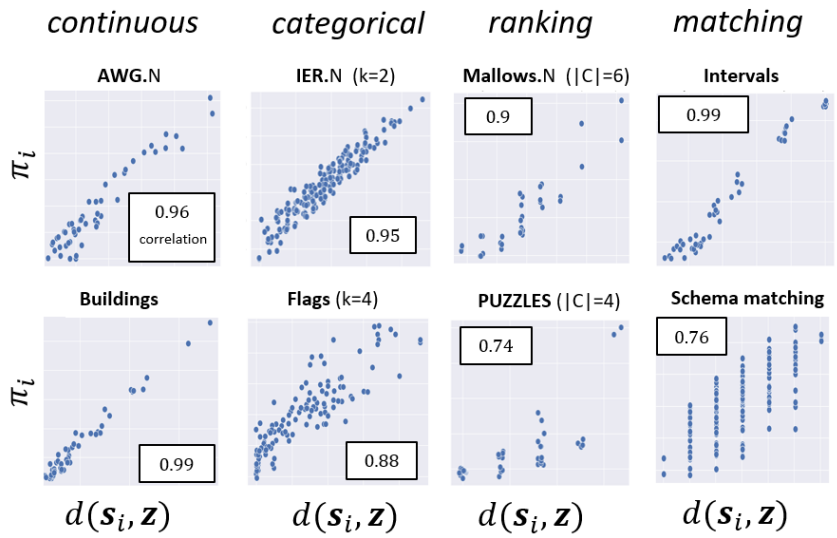

Figure 1: Each scatterplot presents all workers in a single instance, with their disparity π_i vs. their real error $d(\boldsymbol{s}_i, \boldsymbol{z})$.

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.

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 $\min\{n, m\}$.

Noise-independent bounds

What can be said without any assumptions on the distribution of labels? We argue that we can at least tell particularly poor workers from good workers.

Here we show a bound for *binary labels* (that also applies to rankings). In the AAAI23 version we provide a similar result for real-valued labels.

Corollary 2 (Anna Karenina principle for binary labels and rankings)**.** *Assume labels are binary vectors with Hamming distance, or rankings with Kendall-tau distance. For any worker i, $E[\pi_i|t_i] \in [f_i - \mu, f_i + \mu]$.*

Proof. By Theorem 1, $E[\pi_i|t_i] = f_i + \mu_{\mathcal{T}} - 2\langle \boldsymbol{b}_i, \boldsymbol{b}_{\mathcal{T}} \rangle$. When labels are binary vectors, $\langle \boldsymbol{b}_i, \boldsymbol{b}_{i'} \rangle$ simply counts in how many entries $j \leq m$ both $b_{ij} = s_{ij} - z_j$ and $b_{i'j} = s_{i'j} - z_j$ are nonzero, and thus $\langle \boldsymbol{b}_i, \boldsymbol{b}_{i'} \rangle = \min\{f_i, f_{i'}\}$, and

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_{\mathcal{T}} \rangle = E_{i'}[\langle \boldsymbol{b}_i, \boldsymbol{b}_{i'} \rangle] = E_{i'}[\min\{f_i, f_{i'}\}] \leq \min\{f_i, E[f_{i'}]\} = \min\{f_i, \mu_{\mathcal{T}}\}.$$

The upper bound is since $\langle \boldsymbol{b}_i, \boldsymbol{b}_{\mathcal{T}} \rangle \geq 0$, and the lower bound is since

$$E[\pi_i|t_i] = f_i + \mu_{\mathcal{T}} - 2\langle \boldsymbol{b}_i, \boldsymbol{b}_{\mathcal{T}} \rangle \geq f_i + \mu_{\mathcal{T}} - 2\mu = f_i - \mu,$$

as required. □

In particular, this means that given enough samples to accurately estimate $E[\pi]$, we can always distinguish good workers with $f_{i^*} < \mu$ from bad workers with $f_{i^{**}} > 3\mu$, as $E[\pi_{i^*}] < 2\mu < E[\pi_{i^{**}}]$.

Without further assumptions, this condition is tight (i.e. it may not be possible to separate between workers whose faults are μ and 3μ). Consider the following types, for $|C| = 3$:

Meaning	notation	ranking	as binary vector	error f_i
Truth	$\boldsymbol{z}$	abc	000	0
Poor worker	$\boldsymbol{s}_{i^*}$	bac	100	$1/3$
Good worker	$\boldsymbol{s}_{i^{**}}$	cba	111	$3/3$
Common type	$\boldsymbol{s}'$	acb	001	$1/3$

If the mass of the common type is ≈ 1, then $\mu \approx f' = 1$. This means $f_{i^*} \approx \mu$ and $f_{i^{**}} \approx 3\mu$. Yet $\pi_{i^*} \approx d(\boldsymbol{s}_{i^*}, \boldsymbol{s}') = 2/3 = d(\boldsymbol{s}_{i^{**}}, \boldsymbol{s}') \approx \pi_{i^{**}}$ (in fact $\pi_{i^*} = \pi_{i^{**}}$ exactly), so the good and poor workers cannot be distinguished. This example shows that in order to identify the poor workers, we need a reasonable amount of good workers.

Arbitrary Labels and Symmetric Noise

A trivial implication of Theorem 1 is when the average worker is unbiased:

Corollary 3 (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 have arbitrary labels, and 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 4 (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 4 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

Real-valued labels and Gaussian noise One of the most common noise models for real-valued labels is that each worker adds independent Gaussian noise to each label, with mean zero and variance that reflects her fault f_i. This model is known as *Additive White Gaussian* noise. When workers' variance is known, the optimal aggregation is a weighted average, where individual weights are inversely proportional to the variance [1].

An intriguing property of our suggested average similarity method, is that under a slight variation of AWG, it extracts the maximum likelihood estimator for f_i, with a particular regularization factor. This is the main technical result of the AAAI'23 version, and is omitted from this version.

Binary labels Kurvers *et al.* [19] 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 [19] 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 $\mathbb{R}^d$). Note that if we define the type t_i as the probability of providing a false answer (i.e. as $1-p_i$), then the fault $f_i = E[d(\boldsymbol{s}_i, \boldsymbol{z})]$ is exactly t_i.

Corollary 5 (Anna Karenina principle for the binary One-Coin model)**.** *For the binary domain with the One-Coin model,*

$$E[\pi_i | t_i] = \mu_{\mathcal{T}} + (1 - 2\mu_{\mathcal{T}}) f_i.$$

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 [18]) 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.

Rankings According to the *Independent Condorcet Noise* ICN model [39], a worker of type $t_i = f_i^* \in [0,1]$ observes a vector $\vec{s}_i \in \{0,1\}^m$ where for every pair of candidates $j = (a,b)$, we have $s_{ij} \neq z_j$ with probability f_i^*.

By definition, the ICN model is a special case of the binary IER model, where $m = \binom{|C|}{2}$, and the fault level of a type t_i voter is $f_i = f_i^*$. We thus get the following result as an immediate corollary of Corollary 5.

Corollary 6 (Anna Karenina principle for Condorcet Noise model)**.** *For the ranking domain with the independent Condorcet noise model,* $E[\pi_i] = \mu + (1 - 2\mu) f_i$.

4 Aggregation

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

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 [18, 19].

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}. See full version for details on various domain-specific implementations.

while weights can be set heuristically based on (estimated) fault, in some domains this can be guided by theory. In particular, when faults $\vec{f}$ are known then Aitken [1] showed the optimal weights under Gaussian noise are $w_i := 1/f_i$; and Grofman et al. [14] showed that for binary labels setting $w_i^* := \log \frac{1-f_i}{f_i}$ (henceforth 'Grofman weights') is optimal.[4]

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.

[4] In fact, this goes back to Neyman and Pearon [27]. We thank Aryeh Kontorovich for this reference.

ALGORITHM 1: (P-TD) FOR RANKING DATA

Input: Dataset $S \in \mathcal{L}(k)^n$, voting rule r.
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 $\hat{f}_i \leftarrow \pi_i - \bar{\mu}$; `// Step 2`
    Set $w_i \leftarrow \log \frac{1-\hat{f}_i}{\hat{f}_i}$;
end
Set $\hat{\boldsymbol{z}} \leftarrow r(S, \vec{w})$; `// Step 3`
return $(\ddot{f}, \hat{\boldsymbol{z}})$;

Rank aggregation

Note that while our results on fault estimation from the binary domain directly apply (at least to the ICN model), aggregation is more tricky: an issue-by-issue aggregation may result in a non-transitive (thus invalid) solution, and a voting rule is guaranteed to output a valid ranking. While there is much literature on which voting rule comes closest to the ground truth (see, e.g., [3, 11]), we assume here that the voting rule has already been fixed, and are concerned mainly about the effect of weighing voters according to their estimated fault.

The Kemeny rule and optimal aggregation It is well known that for both Condorcet noise model and Mallows model (a restriction of Condorcet model to transitive orders), when all voters have the same fault level, the maximum likelihood estimator of $L_{\boldsymbol{z}}$ is obtained by applying the *Kemeny-Young* (henceforth, *KY*) voting rule on S, see Young [39].

The KY rule r^{KY} computes the binary vector $\boldsymbol{y}^0$ that corresponds to the majority applied separately on every pair of candidates (that is, $\boldsymbol{y}^0 := \text{sign} \sum_{i \leq n} \boldsymbol{s}_i$); then $r^{KY}(S) := \text{argmin}_{L \in \mathcal{L}(C)} d_H(\boldsymbol{x}_L, \boldsymbol{y}^0)$.

A natural conjecture is that applying a *weighted version* of KY with Grofman weights $\vec{w}^*$ that is an MLE when fault levels f_i are known. We did not find any explicit reference to this question or to the case of distinct fault levels in general.[5] As it turns out, the conjecture is correct. This result may be of independent interest. The full proof is in full version.

Theorem 7. *Suppose that the ground truth $L_{\boldsymbol{z}}$ is sampled from a uniform prior on $\mathcal{L}(C)$. Suppose that instance $I = \langle S, \vec{z} \rangle$ is sampled from population $(f_i)_{i \in N}$ via the ICN model where f_i are known. Then $\hat{L} := r^{KY}(S, w^*(f_1), \ldots, w^*(f_n))$ is the MLE of $L_{\boldsymbol{z}}$.*

Proof sketch. Recall that the original proof for weighted majority [14] divided the profile space into two equal and symmetric parts, according to whether they are classified as positive or negative. We proceed in a similar fashion, carefully combining ingredients from the proofs in [14] and in [39]: the profile space is now partitioned into $|C|!$ parts instead of 2, one for each potential ground truth ranking L'. We explicitly compare the log-likelihood ratio of $\hat{L}$ and L' to show that the likelihood of the selected ranking is (weakly) higher than the likelihood of any other order. □

[5] There are some other extensions: [11] deal with a different variation of the noise model where it is less likely to swap pairs that are further apart; and in [36] the KY rule is extended to deal with more general noise models and partial orders.

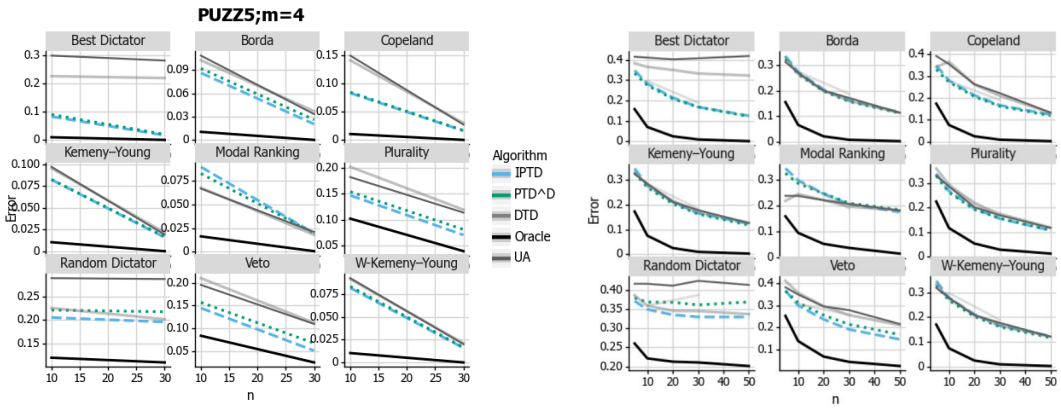

Figure 2: Error on PUZZ5 (Left) and DOTS3 (Right) under nine different voting rules.

5 Empirical Evaluation

In the AAAI23 version, we compared our P-TD and IP-TD algorithms to many other truth discovery algorithms on various datasets from multiple domains. Here we focus on the results for aggregating rankings. Recall that the underlying voting rule is assumed to be part of the input and thus we compare under different voting rules.

While many truth-discovery methods (some estimate workers' competence) have been suggested in the literature, almost all of them are dedicated either to binary labels or to real-valued labels. We therefore consider a simple baseline of *unweighted aggregation* (UA, that simply applies the selected voting rules without weights), and three other algorithms that are domain-independent: MAS [4], TOP2 and EXP [17]. Another possible baseline is to first use the given voting rule for aggregation, then evaluate workers' by their distance from the result and aggregate again (D-TD).

Finally, Oracle Aggregation (OA) is a benchmark using the actual fault f_i of each worker.

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

DOTS subjective rankings of four images of dots, according to the number of dots they contain.

PUZZ contain subjective rankings of four images 8-puzzle boards, according to the number of steps from solution.

BUILD a dataset we collected by asking subjects to evaluate the height of buildings in pictures, and extracting the rankings.

DOTS and PUZZ each contain four datasets of tasks with different difficulty. Both datasets are from Mao et al. [25].

For aggregation, we used nine different ordinal voting rules, see full version for details.

In addition we generated synthetic datasets from Mallows model: in SYS.N workers parameter (accuracy) was sampled from a truncated Normal distribution, and in SYS.HS there were 20% accurate workers ('Hammers') and 80% random workers ('Spammers').

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.

Ranking	v. rule	EXP	TOP2	MAS	P-TD^D	IP-TD
SYN.HS	Borda	-5%	**-24%**	-14%	-4%	-10%
SYN.N	Borda	-3%	+6%	+1%	**-4%**	**-6%**
BUILD	Borda	+0%	+0%	+0%	**-0%**	+0%
DOTS3	Borda	+2%	+7%	+2%	**-1%**	+1%
DOTS5	Borda	+1%	+9%	+4%	**-1%**	**-1%**
DOTS7	Borda	+1%	+13%	+5%	**-3%**	**-2%**
DOTS9	Borda	-3%	+7%	-1%	-6%	**-10%**
PUZZ5	Borda	+1%	+1%	+8%	**-2%**	**-2%**
PUZZ7	Borda	-10%	-7%	-18%	-15%	**-20%**
PUZZ9	Borda	+6%	+48%	+14%	**-5%**	-1%
PUZZ11	Borda	-0%	+6%	+3%	**-3%**	**-3%**
SYN.N	Plurality	**-4%**	-2%	**-5%**	**-4%**	**-5%**
BUILD	Plurality	-1%	+20%	-2%	**-6%**	**-6%**
DOTS3	Plurality	+2%	+14%	+2%	**-2%**	**-1%**
PUZZ5	Plurality	+1%	+12%	+1%	**-3%**	**-2%**
SYN.N	Copeland	-12%	**-37%**	-25%	-27%	-29%
BUILD	Copeland	-12%	-7%	-17%	**-27%**	**-27%**
DOTS3	Copeland	+0%	+0%	**-1%**	**-0%**	**-0%**
PUZZ5	Copeland	-7%	-12%	-16%	**-19%**	**-18%**

Table 2: Results (RI) for rankings datasets, under three different voting rules ($n = 10$, four ranked alternatives), and on the other complex annotation datasets.

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

We generated populations from Mallows distribution, where the proto population (distribution of ϕ_i) was either a clipped Normal distribution (we used $N(0.65, 0.15)$ and $N(0.85, 0.15)$), or a "hammer-spammer" (HS) distribution (20% hammers with $\phi_i = 0.3$, and 80% spammers with $\phi_i = 0.99$ which is just slightly better than random).

We can see in Table 2 that P-TD yields substantial improvement over unweighted aggregation, and beats the other, more sophisticated algorithms on most datasets. Note that while these algorithms are sometimes best (e.g. TOP2 on some synthetic datasets) they are highly unstable and often perform worse than UA.

Fig. 2 shows how P-TD, IP-TD and D-TD improve as the dataset grows, under all nine voting rules. Again we can see that P-TD and IP-TD are better, especially when using simple voting rules: when using KY this is less substantial.

Finally, Fig.3 compares P-TD to D-TD (on synthetic data), showing improvement for all voting rules and scales.

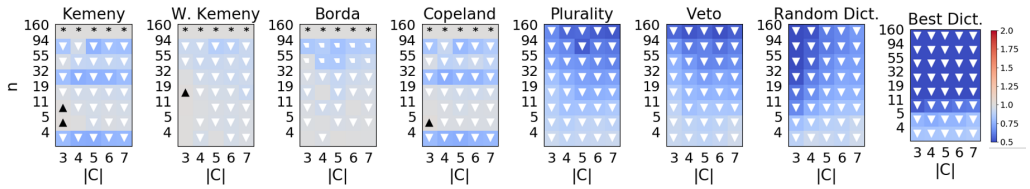

Figure 3: Performance comparison of P-TD to D-TD, with all eight voting rules. The proto-population is Mallows distribution with $\phi_i \sim N(0.85, 0.15)$. A white down pointing triangle indicates an advantage to P-TD.

6 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 [8, 16, 21]; semi-supervised learning [38]; or worker's competence that varies across task types [4].

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.

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

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

[2] Bahadir Ismail Aydin, Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, and Murat Demirbas. Crowdsourcing for multiple-choice question answering. In *Proceedings of the 26th IAAI Conference*, 2014.

[3] Ruth Ben-Yashar and Jacob Paroush. Optimal decision rules for fixed-size committees in polychotomous choice situations. *Social Choice and Welfare*, 18(4):737–746, 2001.

[4] Alexander Braylan and Matthew Lease. Modeling and aggregation of complex annotations via annotation distances. In *Proceedings of The Web Conference 2020*, pages 1807–1818, 2020.

[5] Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? In *Proceedings of the 14' Conference on Economics and Computation (EC'13)*, pages 143–160. ACM, 2013.

[6] Randy L Carter, Robin Morris, and Roger K Blashfield. On the partitioning of squared euclidean distance and its applications in cluster analysis. *Psychometrika*, 54(1):9–23, 1989.

[7] Marie J Condorcet. 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., 1785.

[8] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the Web Conference (WWW'13)*, pages 285–294, 2013.

[9] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28, 1979.

[10] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. Scalable multi-label annotation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'14)*, pages 3099–3102. ACM, 2014.

[11] Mohamed Drissi-Bakhkhat and Michel Truchon. Maximum likelihood approach to vote aggregation with variable probabilities. *Social Choice and Welfare*, 23(2):161–185, 2004.

[12] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction*, 24(4):1–26, 2017.

[13] Chao Gao and Dengyong Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*, 2013.

[14] Bernard Grofman, Guillermo Owen, and Scott L Feld. Thirteen theorems in search of the truth. *Theory and Decision*, 15(3):261–278, 1983.

[15] Shih-Wen Huang and Wai-Tat Fu. 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)*, pages 639–648, 2013.

[16] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'11)*, pages 1953–1961, 2011.

[17] Yasushi Kawase, Yuko Kuroki, and Atsushi Miyauchi. Graph mining meets crowdsourcing: Extracting experts for answer aggregation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'19)*, 2019.

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

[19] Ralf HJM Kurvers, Stefan M Herzog, Ralph Hertwig, Jens Krause, Mehdi Moussaid, Giuseppe Argenziano, Iris Zalaudek, Patty A Carney, and Max Wolf. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science advances*, 5(11):eaaw9011, 2019.

[20] Jiyi Li, Yukino Baba, and Hisashi Kashima. Incorporating worker similarity for label aggregation in crowdsourcing. In *International Conference on Artificial Neural Networks*, pages 596–606. Springer, 2018.

[21] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4):425–436, 2014.

[22] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: is the problem solved? *Proceedings of the VLDB Endowment*, 6(2):97–108, 2012.

[23] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

[24] Tyler Lu and Craig Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *Journal of Machine Learning Research*, 15(1):3783–3829, 2014.

[25] Andrew Mao, Ariel D Procaccia, and Yiling Chen. Better human computation through principled voting. In *Proceedings of the Conference on Artificial Intelligence (AAAI'13)*, 2013.

[26] Elliot McLaughlin. Image overload: Help us sort it all out, nasa requests. CNN.com. Retrieved at 18/9/2014, 2014.

[27] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[28] Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *PNAS*, 111(4):1253–1258, 2014.

[29] Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532, 2017.

[30] Ariel D Procaccia, Nisarg Shah, and Yair Zick. Voting rules as error-correcting codes. *Artificial Intelligence*, 231:1–16, 2016.

[31] Helen Spiers, Harry Songhurst, Luke Nightingale, Joost De Folter, Zooniverse Volunteer Community, Roger Hutchings, Christopher J Peddie, Anne Weston, Amy Strange, Steve Hindmarsh, et al. Deep learning for automatic segmentation of the nuclear envelope in electron microscopy data, trained with volunteer segmentations. *Traffic*, 22(7):240–253, 2021.

[32] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8): 58–67, 2008.

[33] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *ACM SIGIR Workshop on CIR' 11*, pages 21–26, 2011.

[34] Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. Towards building a high-quality workforce with mechanical turk. *NeurIPS workshop*, pages 1–5, 2010.

[35] Jens Witkowski, Yoram Bachrach, Peter Key, and David Parkes. Dwelling on the negative: Incentivizing effort in peer prediction. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, 2013.

[36] Lirong Xia and Vincent Conitzer. A maximum likelihood approach towards aggregating partial orders. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'11)*, 2011.

[37] Houping Xiao, Jing Gao, Zhaoran Wang, Shiyu Wang, Lu Su, and Han Liu. A truth discovery approach with theoretical guarantee. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (SIGKDD'16)*, pages 1925–1934, 2016.

[38] Xiaoxin Yin and Wenzhao Tan. Semi-supervised truth discovery. In *Proceedings of the Web Conference (WWW'11)*, pages 217–226, 2011.

[39] H Peyton Young. Condorcet's theory of voting. *American Political science review*, 82(04):1231–1244, 1988.

[40] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1):3537–3580, 2016.

[41] Bo Zhao and Jiawei Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 1817, 2012.