Proceedings of the Fifth Conference on
Human Computation and Crowdsourcing
(HCOMP 2017)

Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing

Alessandro Checco
University of Sheffield, UK
a.checco@sheffield.ac.uk

Kevin Roitero
University of Udine, Italy
roitero.kevin@spes.uniud.it

Eddy Maddalena
University of Southampton, UK
e.maddalena@soton.ac.uk

Stefano Mizzaro
University of Udine, Italy
mizzaro@uniud.it

Gianluca Demartini
University of Queensland, Australia
g.demartini@uq.edu.au

Abstract

In the context of micro-task crowdsourcing, each task is usually performed by several workers. This allows researchers to leverage measures of the agreement among workers on the same task, to estimate the reliability of collected data and to better understand answering behaviors of the participants. While many measures of agreement between annotators have been proposed, they are known for suffering from many problems and abnormalities. In this paper, we identify the main limits of the existing agreement measures in the crowdsourcing context, both by means of toy examples as well as with real-world crowdsourcing data, and propose a novel agreement measure based on probabilistic parameter estimation which overcomes such limits. We validate our new agreement measure and show its flexibility as compared to the existing agreement measures.

1 Introduction

In many research fields, micro-task crowdsourcing has proven to be an effective alternative/supplement to experts and to automated techniques. A practical example is the crowdsourcing of relevance judgments for Information Retrieval (IR) evaluation, where the relevance of a document to a query is assessed by different crowd workers. This has shown to be an effective, low-cost alternative methodology for such a much needed manual activity in IR effectiveness evaluation. It is common to assign the same task to multiple workers and then to aggregate their answers by, for example, majority vote.

Agreement amongst crowd workers is often overlooked, yet it is an important index to assess the properties of the gathered labels (Paritosh 2012), and to compute a final relevance value from a group of judgments (Aroyo and Welty 2014). Another example of popular crowdsourcing tasks are surveys, where asking for ratings over a multiple-level scale (Likert scale) is common practice.

Measuring agreement among raters is not a new problem and many measures have been defined for that. However, all existing agreement measures have been originally designed for controlled experimental settings where few expert assessors are asked to double-judge a small set of items to measure agreement (e.g., in the context of medical diagnoses). The problem of defining a sound and robust measure of agreement is challenging and multi-faceted: many measures have been proposed to address the problems of missing values, correction by chance, weighting judgment scale in ordinal data, etc.

In this work we focus on interval data (like, e.g., relevance scales), and on scenarios where a crowd of undifferentiated and potentially always changing workers (like, e.g., on a paid micro-task crowdsourcing platform) judge a group of items.

The majority of agreement measures is borrowed from data reliability theory, where the reliability of a set of grouped measurements is assessed via a comparison between the inter-group and the intra-group variability, and where typically the judgments are made by a fixed set of assessors. In the context of crowdsourcing, these measures suffer from many problems when used to estimate agreement instead of data reliability:

- The variability of the judgments is typically higher when the judgments concentrate around the center of the scale. This problem is intrinsic to finite scale judgments and can lead to overestimating disagreement over items where the truth concentrates around the scale boundaries.
- The values around which the judgments concentrate (if any) can be different item by item. This can lead to overestimating expected disagreement and thus increasing the possibility of considering the data as random.
- For some items a ground truth (e.g., 'gold questions' in crowdsourcing) might be present, that is a value around which judgments are expected to concentrate. This information is typically not used by classic agreement measures.
- The global variability-based correction by chance leads to many idiosyncrasies in the existing measures, making them hard to use in a crowdsourcing setting.

Our goal in this paper is to address the aforementioned issues, and to build a framework more suitable to estimate worker agreement over a group of tasks in a crowdsourcing context. The main contributions of this paper are:

- We define Φ, a novel measure of agreement for a group of rating tasks.

- We validate Φ against a set of known limitations of existing agreement measures via synthetic scenarios.
- We validate Φ on a real crowdsourcing dataset.
- We provide a Bayesian inferential framework to perform statistical tests on the goodness of agreement.

This paper is structured as follows. Section 2 details the state of the art by recalling existing agreement measures and Section 3 describes the limitations of such measures. In Section 4 we explain our proposal, Φ, a novel agreement measure based on Bayesian parameter estimation. Section 5 shows the validation of our measure both on notable examples as well as on a real dataset. Finally, Section 6 summarizes the paper, lists some current limitations of our measure, and sketches future work.

2 Existing Agreement Measures

Many measures for computing agreement among annotators are available in the literature. A first and straightforward measure is percent agreement, which represents the portion of cases in which the annotators agree, compared to the total number of observations. This measure can be formalized as

$$a_o = \frac{\text{number of cases coders agree (A)}}{\text{total number of targets analyzed (N)}}.$$

It can be trivially extended to more than two raters considering the ratio of pairwise comparisons with agreements over the total number of possible pairwise comparisons, and to multiple items by considering the mean of the per-item percent agreement. In the rest of the paper, we will refer to the *pairwise percent agreement* with a_o. The main issue with a_o is that it does not take into account that some agreement can arise by chance: as a consequence, the number of possible values in the scale heavily affect this measure. Moreover, a_o only copes with nominal data.

Two very similar chance-corrected measures of agreement were developed by Scott (1955) (Scott's π) and Cohen (1960) (Cohen's κ). These measures consider the case of two assessors, and assume a nominal scale for rating the items. The difference between π and κ is that π assumes annotators to provide answers following the same distribution.

Measures like κ are very popular in crowdsourcing experiments to study agreement among workers completing the same task (Aroyo and Welty 2014). However, Feinstein and Cicchetti (1990) and Byrt, Bishop, and Carlin (1993) criticized κ discovering that its value is affected by population variability (i.e, the variability of the data). Furthermore, Hutchinson (1993) claims that the value of κ is affected by two different kinds of disagreement related to the workers' bias (Banerjee et al. 1999).

To consider different "levels" of (dis)agreement (e.g., in a medical domain, a disagreement between two doctors for a diagnosis is more serious than a disagreement between one doctor and a patient), a weighted version of κ has been proposed by Cohen (1968).

A popular alternative to κ is represented by the Intra-class Correlation Coefficient (ICC) proposed by Bloch and Kraemer (1989). This measure considers agreement when workers have the same marginal probability of an answer. To deal with different marginal distributions a modified version of κ has been proposed by O'Connell and Dobson (1984). ICC has proven to be equivalent to (weighted) κ under certain assumptions (Banerjee et al. 1999; Fleiss and Cohen 1973). Furthermore, an alternative version of κ was proposed to deal with the case of multiple assessors (Fleiss 1971), and for the case of ordinal scale of judgment (Cohen 1968; Landis and Koch 1977).

More recently, Krippendorff (2007) attempted to generalize all the previous agreement measures, considering the cases of multiple assessors and of missing data, and dealing with multiple scales (i.e., binary, nominal, ordinal, etc.); the measure, called Krippendorff's α, considers also a measure of "reliability" of the data. The measure can be formalized as

$$\alpha = 1 - \frac{\text{observed disagreement } (D_o)}{\text{expected disagreement } (D_e)}.$$

Here the expected disagreement D_e is used as estimator for the disagreement by chance. We refer to Appendix A for a detailed explanation on the computation of D_o and D_e.

We decided to focus on two measures to use as baselines to compare our novel proposal: Krippendorff's α (as it generalizes and shares the same behavior of previously proposed agreement measures like Cohen's and Fleiss's κ, Intra-class Correlation agreement, Spearman's correlation ρ, and Scott's π) and pairwise percent agreement a_o (because it is complementary to other measures in terms of the limitations it has). Our goal is to propose a novel agreement measure that is able to overcome all the limitations commonly displayed by existing measures when adopted in crowdsourcing contexts.

Notation Given N items, each of them with M judgments, X_{ij} is the j-th rating of the i-th item:

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{d1} & X_{d2} & \dots & X_{dm} \end{bmatrix}.$$

For example, if we have: 4 documents with 5 relevance judgments each, and a scale from 0 to 5, a possible judgment matrix will be:

$$X = \begin{bmatrix} 3 & 2 & 4 & 1 & 0 \\ 0 & 1 & 1 & 5 & 0 \\ 1 & 1 & 0 & 0 & 5 \\ 0 & 2 & 1 & 2 & 1 \end{bmatrix}.$$

3 Limitations of Existing Agreement Measures

We focus now on detailing the limitations of the existing agreement measures. From the exhaustive set of limitations studied by Zhao, Liu, and Deng (2013), we identify the ones that we deem more relevant in the context of crowdsourcing and IR. Since Krippendorff's α has the same behavior, in terms of limitations, of the measures it generalizes (Cohen's and Fleiss's κ, intra-class correlation agreement, Spearman's correlation ρ and Scott's π) we will consider α in the rest

of the paper, together to the somehow complementary (in terms of limitations) percent agreement measure a_o, which we will use as baseline when α leads to unsatisfactory scenarios. It is important to notice that the only limitation a_o suffer from is a very important one (*Random guessing can be reliable/Correction by chance*, as shown in the following), and thus percent agreement cannot be considered as a good candidate for a reliable agreement measure in crowdsourcing and IR. Using the same nomenclature of Zhao, Liu, and Deng (2013), we focus on the following limitations:

1. *Paradox: Random guessing can be reliable / Correction by chance.* Percent agreement (a_o) does not take into account random chance agreement, which leads to a paradox: random answers can lead to obtaining a high agreement score. Let us make an example; imagine a big number of workers (say 100) tackling a very difficult task, using a binary relevance scale (i.e., 0 or 1). The task is so difficult that their answers can be considered random. The percent agreement will be $a_o = 50\%$, which is exactly in the middle between complete agreement and disagreement. On the other hand, if the relevance scale has ten values, the percent agreement will drop to $a_o = 1\%$. For the measures on this paradox in the following we will assume to have 1000 tasks and 100 workers.
2. *Paradox: Random guessing may be more reliable than honest coding/Abnormality: undefined reliability.* According to α, systematic agreement from honest workers can be less reliable than random answers. For example, if we measure a constant (e.g., the gravitational constant, the speed of light, etc.) with perfect precision, we obtain as result of our measure the exact same score (i.e., $a_o = 100\%$ agreement). According to α's definition, the measure is not defined[1], because $D_e = 0$ and $D_o = 0$.
3. *Abnormality: High agreement, low reliability.* When using α, we can have high agreement, but low reliability. Suppose workers have to judge 100 items, using a binary scale; suppose the outcome is: agreement on 99 items (99 positive agreement[2]), and 1 disagreement. a_o would return a 99% of agreement, but $\alpha = 0$ because $D_o = D_e$ (the specific value depending on the number of workers); in fact, α assumes that the data is not reliable at all. This toy-example points out that α mixes together two distinct concepts: assessor agreement and data reliability.
4. *Abnormality: Zero change in a_o causing radical drop in reliability.* Suppose we run an experiment with 100 workers and a binary scale (i.e., 0 and 1); suppose a first outcome (a) of our experiment is 99% agreement (98 positive agreement, 1 negative agreement) and 1% disagreement; in this case we have that $a_o = .99$ and $\alpha = .66$.

   Suppose now a second outcome (b) for our experiment, with the same level of agreement (99%), but with a different distribution: 99 positive agreement, 0 negative agreement, and 1 disagreement; in this case, as before, we obtain $a_o = .99$, but α drops to zero because $D_o = D_e$ (the specific value depending on the number of workers).
5. *Abnormality: Eliminating disagreements does not improve reliability.* We run an experiment with 100 workers and a binary scale (i.e., 0 and 1); suppose a first outcome (a) of our experiment is 90% agreement (90 positive agreement) and 10% disagreement; in this case we have that $a_o = .90$ and $\alpha = -.04$.

   We now move towards perfect agreement (b): suppose 98% agreement (98 positive agreement) and 2% disagreement; in this case we have that $a_o = .98$, but $\alpha = -.05$.

   If we now go from 98% to 99% agreement (c), α becomes 0, but if we reach 100% of agreement (d), α jumps to the value of 1: this discontinuity is clearly not desirable.
6. *Abnormality: Honest work as bad as coin flipping.* Suppose two workers have to identify 50 items in a larger set of 60 items. One worker answers 60 times 1, (i.e., 10 false positives), the other one answers 40 times 1 (i.e., 10 false negative); the outcome (a) thus is 40 agreement (40 positive agreement), and 20 disagreement (20 positive disagreement). In this case $a_o = .66$ and $\alpha = -.19$.

   If workers now answer uniformly at random (b), we get $a_o = 0.5$, and $\alpha \approx 0$ (the value being exactly zero if the realization of the random variable is perfectly uniform), that would seem more reliable than the result ($\alpha = -.19$) obtained from honest workers.
7. *Abnormality: Same quality, same agreement, higher reliability.* Suppose we have to judge 100 items, 50 positive, 50 negative; first outcome (a): 80 agreement (40 of the positive, 40 of the negative) and 20 disagreement (10 of the positive, 10 of the negative). We get $a_o = .8$ and $\alpha = .73$.

   Suppose the outcome where we switch the 40 positive agreement to be negative (b): 80 agreement (0 of the positive, 80 of the negative) and 20 disagreement (10 of the positive, 10 of the negative). We obtain $a_o = .8$ and $\alpha = .26$. We obtained different measure values with the same disagreement level.

   Another example would be to "reverse the question" (e.g, ask "how relevant is this document?" vs. "how NOT relevant is this document?"); in the first case the judgment could be:

$$X = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

   In this case we have that (c) $a_o = .33$ and $\alpha = -.125$. In the second case, when we reverse the outcome, which is equivalent to reverse the question (assuming identical and coherent workers), the outcome (d) of the experiment is:

$$X = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

   In this case, $a_o = .33$, as before, but α changes to .28.
8. *Paradox: Punishing larger sample and replicability.* Suppose two coders, binary scale; 40 items to judge; the outcome (a) is 30 agreement (20 positive agreement, 10

[1] Krippendorff claims that in this case the data should be considered completely unreliable (Kripendorff 2004, p. 425)

[2] If we have a binary scale (i.e., 0 and 1), we refer as positive agreement/disagreement when there is agreement/disagreement around the value of 1, negative agreement/disagreement vice-versa.

Limitation	a_o	α	Φ
1	✗	-	-
2	-	✗	-
3	-	✗	-
4	-	✗	-
5	-	✗	-
6	-	✗	-
7	-	✗	-
8	-	✗	-

Table 1: The limitations considered in this work for a_o and α and Φ marked with "✗" when the limitation is present.

negative agreement) and 10 disagreement. This means $a_o = 0.75$, $\alpha = .47$.

If we make the data ten times larger we get 400 items; 300 agreement (200 positive agreement, 100 negative agreement) and 100 disagreements; this means (b) $a_o = 0.75$, $\alpha = .46$ (lower than before, but no decrease expected).

If we take 4 items, (2 positive agreements, 1 negative agreement) and 1 disagreement, the outcome (c) is a_o= 75%, $\alpha = .53$ (higher than the first case).

Table 1 summarizes the above limitations. The proposed measure Φ behaves correctly in all cases.

4 Our Proposal: Common Agreement Φ

We propose Φ, a measure for agreement of interval data that is able to resolve the paradoxes and abnormalities of the currently used measures, while still being able to perform a substantial level of correction by chance. This new measure is therefore a better option for agreement in a crowdsourcing setting as compared to existing standard agreement measures.

4.1 Intuition

For simplicity, let us first consider a single item with multiple judgments. The intuition behind Φ is connected with the definition of agreement: we consider as *agreement* the amount of concentration around a data value. Conversely, if the data does not concentrate around a value then we have *disagreement* (negative agreement in our measure), that can be more or less strong depending on how polarized the different opinions are. More in detail, our approach can be described as fitting a distribution to the histogram of the judgments and then measuring the dispersion of such distribution. It is important to notice that the fitting distribution has to be general enough to capture the main behaviors that might occur: flat (random judgments), bell-shaped (agreement), J-shaped (agreement around a value on the boundary of the scale), and U shaped distribution (disagreement), as shown in Figure 1. At the same time, the desired distribution has to have a minimal number of parameters, to avoid overfitting. For this reason, we follow the approach of (Smithson and Verkuilen 2006) and use a Beta distribution to perform the fit: Φ is a transformed parameter of the Beta distribution over the histogram of the collected answers. Such parameter is related to the standard deviation of the fitted distribution, with the difference that here we account for the finiteness of the rating scale, and thus we adjust for the tendency of having lower dispersion when the data concentrates around a value on the boundaries of the rating scale. For example, if we imagine a scenario where assessors add a random Gaussian noise to the ground truth when making a judgment, we can immediately see that the dispersion will be minimum when the ground truth is on the boundary of the scale, because a Gaussian noise that would result is a judgment outside the boundary would be clipped.

Multiple Items The strength of our approach becomes apparent when applied to a group of items to be judged: in the case of relevance judgment tasks, each item i is allowed to have a different average relevance value μ_i, while the agreement among workers is defined as the *common* Φ that better explains the judgment data. This allows to solve the problems that arise, in the other agreement measures, when trying to correct by chance by using the dispersion of the whole dataset as normalizing factor: this is the key to solve limitations 2-8 of Table 1.

4.2 Beta Distribution

In order to formalize the definition of Φ, let us start by considering a single item i, judged by M workers. We assume the judgments to be continuous on a finite scale: for simplicity, we will consider the scale equal to $[0, 1]$, simply achievable by a scaling of the data. Discrete cases are considered as a discretization of a continuous judgment intent. Given a set of judgments, we might be tempted to use the standard deviation (std) as measure of disagreement. But due to the nature of bounded distributions, the dispersion of the data depends also on the position of the mean of the data: as above mentioned, the dispersion will be lower on the boundaries than on the middle of the scale. Capturing the case of multiple documents with same data dispersion around different values would become problematic.

We can model quite a large set of distribution types using a Beta distribution $B(a, b)$. We can re-parametrize the distribution in terms of the mean value μ and the precision p as in (Smithson and Verkuilen 2006), by setting:

$$\mu = \frac{a}{a+b}, \quad p = a + b.$$

This parametrization allows us to analyze separately the mean value and the dispersion: when μ is fixed, the dispersion decreases as p increases. This allows to have a measure of dispersion that is agnostic to the position of the mean, solving the common problem of different dispersion for values close to the boundaries of the scale. Finally, we perform a transformation that allows our measure to be comparable with the existing measures by having the image in the interval $[-1, 1]$, with 1 corresponding to maximum agreement, -1 to maximum disagreement, and 0 corresponding to uniform at random response:

$$\Phi(p) = 1 - 2^{\frac{-p \log 2}{2}},$$

and therefore we will refer to the Beta distribution $B(\mu, \Phi)$. Considering μ fixed, $B(\mu, \Phi)$ can capture 4 major agreement behaviors which are shown in Figure 1:

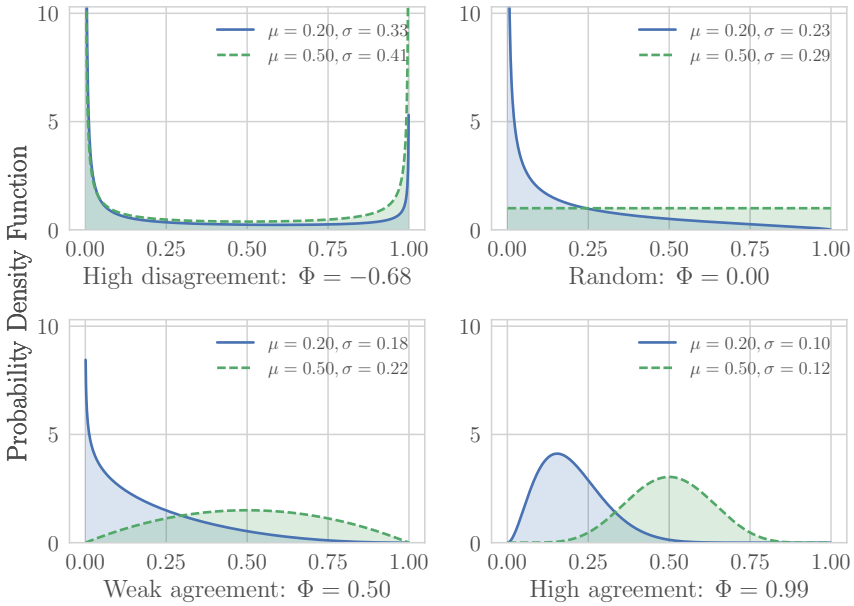

Figure 1: Example of two Beta distributions with mean $\mu = 0.5$ and 0.2 for four different values of agreement Φ.

High Disagreement. When $\Phi < 0$, there is no central tendency value but rather a tendency to exclude a central area (polarized behavior).

Random. When $\Phi = 0$, the behavior is equivalent with a unbounded uniform process censored on the scale.

Weak Agreement. When $0 < \Phi \leq 0.5$, the distribution has no inflection point, but there is a unique central tendency or a dispersion that is smaller than a uniform process.

High Agreement. When $\Phi > 0.5$, the distribution is bell shaped with two inflection points, more narrow around the mean as Φ grows.

To summarize, in this context the Beta distribution is the natural choice to model interval data in a bounded interval in a general enough way without the expenses of many parameters, that would lead to overfitting. Moreover, the choice to reparametrize in terms of μ and Φ allows us to model the agreement in a way that is agnostic from the position of the mean data value μ.

4.3 Model Assumptions

We develop now a simple inferential Bayesian[3] framework to compute the agreement Φ for a group of items $i = 1, \ldots, N$ and workers $j = 1, \ldots, M$[4] . We make the following assumptions:

Assumption 1. Worker j judges item i according to a $B(\mu_i, \Phi)$ distribution. This assumption is able to encompass behaviors that are random, polarized or aggregated around a mean μ_i as shown in Figure 1.

Assumption 2. Workers are indistinguishable and their behavior is independent from each other and from past judgments.

Assumption 3. Items $i = 1, \ldots, N$ share a common sense of *agreement*, in the sense that regardless on where the central tendency (or even the complete polarization) is on a specific item, the value of Φ will be shared by all workers for all items.

While assumptions 1 and 2 are somehow standard, Assumption 3 is the main fulcrum of the model: it allows a parsimonious parametrization and aims to capture the common agreement or disagreement properties (if present) in the group of items. For a more detailed analysis on this assumption and its consequences see Section 4.6, where we show a way to verify the goodness of this assumption.

Given these assumptions and a set of observed values $X_{i,j}$, we obtain the following log-likelihood (i.e., the conditional distribution over the observed judgments):

$$\mathrm{P}(X|\vec{\mu}, \Phi) \propto \prod_{i=1}^{N} \prod_{j=1}^{M} B(X_{i,j}|\mu_i, \Phi)^{O_{ij}},$$

where O_{ij} is equal to one if worker j rated item i, and zero otherwise. We assign the following prior distributions on the parameters:

$$\mathrm{P}(\vec{\mu}|\sigma_\mu^2) = \prod_{i=1}^{N} \mathcal{N}(1/2, \sigma_\mu^2 \mathbf{I})$$
$$\mathrm{P}(\Phi|\sigma_\Phi^2) = \mathcal{N}(0, \sigma_\Phi^2),$$

[3]We will discuss the corresponding non Bayesian approach in Section 4.4.

[4]The number of workers can be different for each item, as explained later.

where $\mathbf{I}$ is the identity matrix. If we do not have additional information, we set the hyperparameters $\sigma_\mu^2 = \sigma_\Phi^2 \to \infty$, to express our initial absence of knowledge about the parameters. For numerical reasons, it is suggested to simply set the hyperparameters to a big number. In Section 4.7 we discuss how to incorporate additional information about the items (i.e., possibly available ground truth) in the model.

The predictive posterior distribution over Φ and $\vec{\mu}$, given the observed data X, O is then:

$$\mathrm{P}(\vec{\mu}, \Phi | X) = \prod_{i=1}^{N} \prod_{j=1}^{M} B(X_{i,j} | \mu_i, \Phi)^{O_{ij}} \prod_{i=1}^{N} \mathcal{N}(1/2, \sigma_\mu^2 \mathbf{I}) \mathcal{N}(0, \sigma_\Phi^2) C, \quad (1)$$

where C is a normalization constant that does not depend on the parameters.

4.4 MAP Estimation of Φ

To quickly obtain a simple estimator of Φ, we can now maximize Equation (1) to find the Maximum A posteriori Probability (MAP) estimate of the parameters. The measure of agreement for a group of items is then

$$\hat{\Phi} = \arg\max_{\Phi} \mathrm{P}(\vec{\mu}, \Phi | X).$$

This operation is computationally trivial in modern computers, but does not allow to understand *the confidence* of the estimation. This is why we recommend to sample from the posterior distribution to obtain a robust estimation of Φ and a confidence interval, as shown in Section 4.5.

Non Bayesian Approach. Taking the limit of the hyperparameters to ∞, the MAP estimates becomes equivalent to a Maximum Likelihood Estimation, but we believe that the Bayesian machinery of the next section is able to significantly improve the usability of our model. Moreover, a Bayesian approach allows to easily incorporate information about the ground truth, as explained in Section 4.7

4.5 Sampling from the Posterior Distribution

We can obtain an estimate of the distribution of the parameters by sampling from the log-posterior $\mathrm{P}(\vec{\mu}, \Phi | X)$, and a credible interval (akin to a confidence interval) to signal the confidence of the agreement measure Φ.

In Figure 2 we show a representation of the inference results for the judgments of 17 documents. We generated a small synthetic dataset, where the first document has an outlier on the right boundary, and the other 16 documents have a clear central agreement. In the figure it can be see that documents 2-5 are replicated four times to get 16 documents that have higher agreement. We can see that the model is forced to find the best agreement level (dispersion of the Beta distribution) that collectively explain all the data: while document 1 alone would have been fitted with a high disagreement (a U shaped) Beta, the most probable Beta for the model to explain the whole dataset is the one where the first document

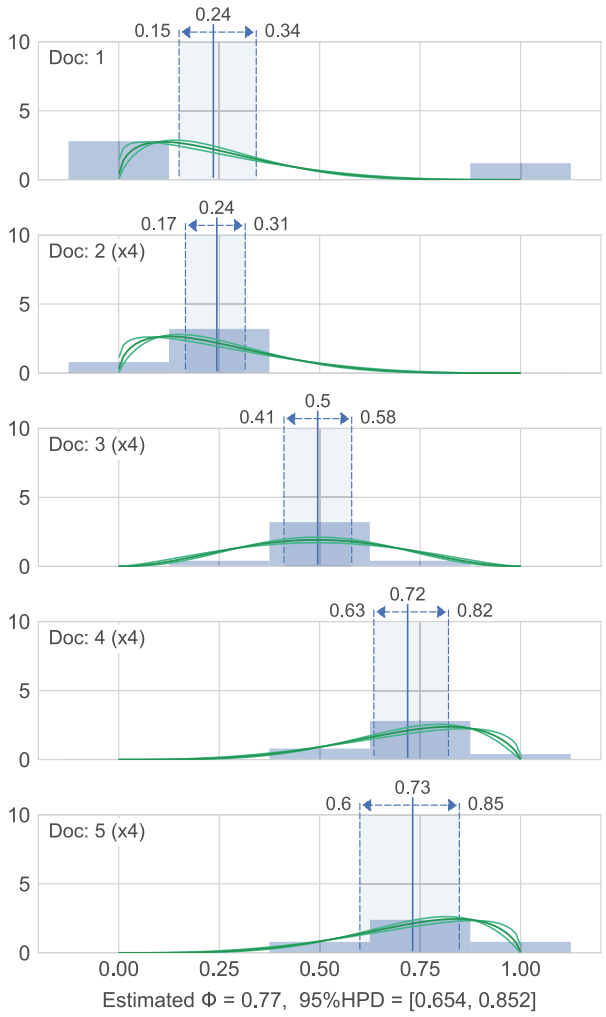

Figure 2: Inference result for a group of 17 documents (documents 2-5 are replicated four times). The thick green line is the estimated Beta for each document, with the 95% HPD interval curves surrounding the best estimate. The overall agreement is high (0.77).

has an outlier. This reflects the way we perceive the agreement level as humans, especially with a small set of data samples, and allows to get a robust estimation of agreement for group of documents.

Once the log-posterior distribution has been estimated, we will be provided with an estimate of $\vec{\mu}$ and Φ and their 95% Highest Posterior Density (HPD) intervals (the shaded regions around the mean for μ and the thin lines around the fit for Φ in Figure 2).

4.6 HPD Interval and Significance Testing

Similarly to a frequentist statistical test, once we have an estimate of the 95% HPD interval of Φ, we can claim to have agreement (resp. disagreement) if the HPD interval is positive (resp. negative) and does not contain 0. See Section 5.2 for an analysis on the bias of the statistical test.

If the HPD interval is very large, for example including positive and negative values, it would mean that a measure of agreement for the dataset cannot be reached, either because a larger sample size is needed, or because the documents in the dataset are very different in term of agreement: a dataset clustering might be needed.

4.7 Incorporating Ground Truth

It is also possible to incorporate knowledge of the ground truth g_i for item i by simply setting up the prior:

$$\mathrm{P}(\mu_i|\sigma_\mu^2) = \prod_{i=1}^{N} \mathcal{N}(g_i, \sigma_\mu^2),$$

with σ_μ^2 equal to a relatively small number (proportional to the degree of certainty on the fact that the workers will aggregate around that value). This will allow to better discriminate outliers for that specific item.

A practical example is represented by the task of crowdsourcing relevance judgments of documents; we can use as a ground truth the actual label (i.e., relevance value) of the document. In this case the prior tells to the model the value around which we expect the data should aggregate.

5 Validation

While our measure Φ has its ideal application in interval data and crowdsourced data, we first validate it on the situations (typically described in an IR context) where other agreement measures are unsatisfactory.

5.1 Validation on Notable Examples

In this section we validate Φ against α and a_o on the notable examples where currently used measure prove to be unsatisfactory, as explained in detail in Section 3. Table 2 shows the results. In the left column we refer to the examples where the paradoxes and abnormalities described in Table 1 apply. We compare the ability of the three measures to cope with the paradoxes.

Results show that Φ is not affected by any of the paradoxes considered, maintaining the same good behavior of a_o, without the problems of Paradox 1 and the lack of correction by chance (see Table 1).

5.2 Bias Analysis

Our model does not make the assumption that every item in a group has the same μ. In this way the model manages to capture agreement in cases where other measures fail to distinguish the data from the uniform distribution. This capability comes at a cost: the increased statical power introduces biases which we discuss next.

Number of Workers. When a small number of workers is present, the inherent nature of subsampling from a uniform distribution will lead to overestimate the agreement for a uniform at random behavior. That means that if a small set of workers answer uniformly at random in the scale, we will not be able to detect that by our method (type I error). In Figure 3 an analysis of the bias of the estimator shows

Limitation	a_o	α	Φ	95% HPD
1	(*)	0.005	0.05	[.045,.053]
2	1	NaN	1	[1, 1]
3	0.99	0	1	[1, 1]
4a	0.99	0.66	1	[1, 1]
4b	0.98	0	1	[1, 1]
5a	0.90	−0.04	0.99	[.99, .99]
5b	0.98	−0.05	1	[1, 1]
5c	0.99	0	1	[1, 1]
5d	1	1	1	[1, 1]
6a	0.66	−0.19	0.89	[.69, .97]
6b	0.50	0	0.73	[.52, .88]
7a	0.8	0.73	0.7	[.6, .8]
7b	0.8	0.26	0.7	[.6, .8]
7c	0.33	−0.125	0.15	[-.4, .5]
7d	0.33	0.28	0.15	[-.32, .5]
8a	0.75	0.47	0.95	[.84, .99]
8b	0.75	0.46	0.97	[.94, .98]
8c	0.75	0.53	0.96	[.14, .99]

Table 2: Validation of Φ, with HPD intervals, using the examples shown in Section 3. Regarding limitation 1 (*), a_o can assume any value from 50% downwards, depending on the number of discrete values in the scale.

that if, e.g., only 5 workers are used, there is a bias of about 0.5. This means that it is preferable to only consider strong agreement ($\Phi > 0.5$) as indication of agreement.

More work on the analysis of bias and correction of it is left for future work, but it is important to notice that no statistical test with meaningful power (low type II error) would be able to discern a uniform distribution hypothesis without bias for a small number of workers. In other words, statistical tests that tend to correctly reject a uniform null hypothesis will have low statistical power.

Discrete Values. When the data are discretized or collected over a small number of discrete values, the inherent nature of subsampling from a uniform distribution will lead to overestimate the disagreement for a uniform at random

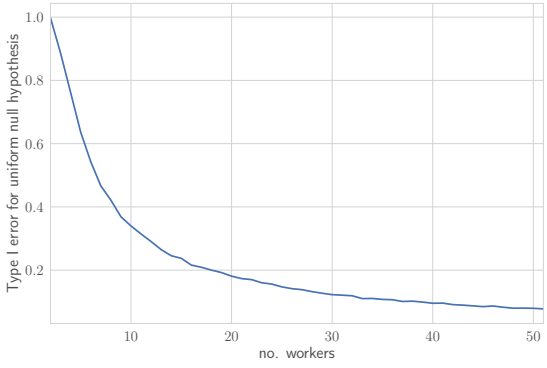

Figure 3: Φ estimator bias against uniform hypothesis.

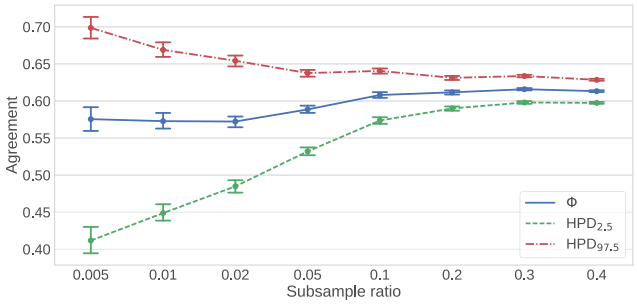

Figure 4: Agreement score vs ratio of the data-frame used for the random sampling; the three curves represent the progress of Φ, together with its $\text{HPD}_{2.5}$ and $\text{HPD}_{97.5}$.

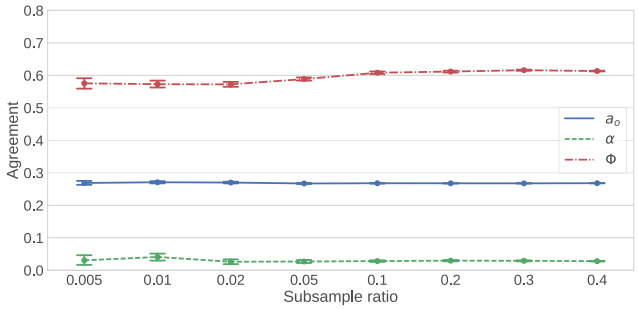

Figure 5: Agreement score vs ratio of the data-frame used for the random sampling; the three curves represent the trend of a_o, α, and Φ.

behavior. That means that if a small group of workers answers uniformly at random in the scale, our method will be biased toward disagreement. From an analysis of the bias analogous to the one carried out in the previous section, we can conclude that this method is very robust when a scale of at least 5 points is provided.

5.3 Performance Estimation

Using the python PyMC3 library (Salvatier, Wiecki, and Fonnesbeck 2016), the estimation of Φ and its 95% HPD can be computer in under a minute on 5 Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz cores for a group of 7000 items and 5 workers. To obtain a similar result (with a confidence interval) for Krippendorff's alpha it would be necessary to run a bootstrapping on the whole dataset, that would require a time of 3 orders of magnitude longer.

5.4 Validation on Real Data

To validate our measure experimentally we use the publicly available dataset produced in (Gadiraju, Yang, and Bozzon 2017), comparing the agreement measures a_o, α, and Φ.

The data consists of the evaluation of the clarity of the instructions of a set of 7100 crowdsourcing tasks (belonging to 6 distinct task types, as categorized in (Gadiraju, Kawase, and Dietze 2014)), where each task has been judged by 5 undifferentiated workers. There is no expectation for tasks of different types to have the same average clarity score, so our Φ measure seems appropriate to estimate the agreement of the whole set.

We check whether the measure is stable w.r.t to addition of data, and how the HPD evolve. Figure 4 shows the evolution of Φ varying the fraction of the dataset considered (100 random subsample per point). The plot shows the estimated Φ, $\text{HPD}_{2.5}$ and $\text{HPD}_{97.5}$. We can see that as the amount of data available increases, the HPD interval shrinks. Following the recommendations of Section 5.2, we are able to claim that with 95% confidence there is significant positive agreement in the data. Figure 5 shows a_o, α and Φ (HPC interval omitted) varying the fraction of the dataset considered in Figure 4. We can see that the three measures are very different; a_o is constant, no matter which sample ratio we consider; α's value drops as the sample ratio grows; this means the more items we consider, the more α consider the agreement to be random; Φ, contrary to α, increases its value the more sample ration we consider; this means that for Φ, in this dataset, some agreement arises if we increase the percentage of items we consider.

6 Discussion and Future Work

6.1 Summary

In this paper, we discussed the limitations of the existing agreement measures, which cannot deal with certain scenarios frequently occurring in crowdsourcing settings. Thus, a new agreement measure is needed. We then introduced Φ, an agreement measure that is able to resolve the paradoxes and abnormalities of the currently used measures, while still being able to perform a substantial level of correction by chance.

Φ is rather general in term of the data it can handle, but is particularly adequate (if not ideal) for subjective tasks, i.e., tasks where the workers can legitimately express rather different judgments and the experimenter cannot expect the different documents' judgments to aggregate around the same mean.

Our experimental validation results show that for our measures to achieve high confidence results, there should be at least 5 workers per task, and the scale used should be continuous or discrete with at least 5 points.

The software implementation of our proposed measure is available open-source at https://github.com/AlessandroChecco/agreement-phi.

6.2 Limitations of Our Measure

Continuous Data Assumption. Our model assumes that workers provide judgments in a continuous scale. If only discrete values are available, the model considers the measurement as a noisy discretization of a real judgment. This can be a problem for very narrow discrete scales (e.g., $\{0, 1\}$). In such cases (i.e., agreement for binary classification tasks) a simpler agreement measure like percent agreement or Cohen's κ may be more appropriate.

Nominal Data. Our model heavily relies on a notion of distance amongst ratings. A relaxation to nominal data would

probably be a stretch, but we leave this analysis for future work.

Polarized Judgment. Our model considers a set of ratings in disagreement when a central tendency cannot be found, and the level of variability is bigger than uniform random judgment. This is modeled with a beta distribution, so the cases in which more than two peaks are present in the judgment distribution might be considered more similar to random selection than disagreement.

Single Item. While our model is well defined even for only one item (especially considering the HPD interval on the estimated agreement), it has high chance of over-fitting the data in this case. This problem is shared with all other measures. The same problem arises when too few items (or workers) are present, but we argue that our model can mitigate this limitation by providing a (potentially rather large) HPD interval when the sample size is too small thus alerting of the low confidence of the agreement measure. In crowdsourcing experiments this is often not an issue as many items and workers are typically considered.

Closed Form. Obtaining a closed form of the MAP and especially of the HPD interval is very difficult: it is usually necessary to numerically compute those with an optimization routine. With modern techniques of sampling, this operation is feasible and not time-consuming.

Bias. While our technique decreases type II errors, it can increase type I errors, especially for a small number of workers and for a small number of possible values when the scale is discretized. This can be expected, because no test with high enough statistical power would be able to discern a uniform distribution hypothesis without bias for a small number of workers/scale values. Conversely, statistical tests that tend to always correctly reject a uniform null hypothesis will have very low statistical power. From a quick study of the bias (Section 5.2), we conclude that having at least 5 workers and 5 values in the scale makes the bias small enough for practical purposes.

6.3 Future Work

The proposed measure Φ solves many of the common issues that arise when assessing agreement in crowdsourcing settings, but is far from perfect. Possible directions of improvement for such techniques include:

- A more detailed analysis of the bias.
- The derivation of a closed form for the MAP.
- A more efficient implementation of the HPD estimator.
- A more extensive testing on real datasets.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732328 and was partially supported by the UK EPSRC grant number EP/N011589/1. We want to thank the Erasmus+ traineeships program for facilitating collaborations, and the European Science Foundation for funding the Science Meeting SM 5917 under the ELIAS Research Networking Programme.

A Appendix - Computing Krippendorff's α

We will follow the notation in (Honour 2016) to summarize how to compute Krippendorff's α.

Let R the set of potential responses that a judgment may yield: R has to be a finite set. Note that in general this information might not be inferred from the data only, because some values might never appear.

The judgments to a given item form a multiset (a set that can have repeated instances) u. Let $\delta(c,k)$ be the (pseudo) distance between judgments c and k, such that $\delta(c,c) = 0$, $\delta(c,k) \leq 0$, and $\delta(c,k) = \delta(k,c)$. This function δ can be defined in different ways, to account for nominal, interval, or ordinal data (Krippendorff 2007). The *average disagreement* within a multiset u is defined as:

$$\mathrm{D}(u) = \sum_{c \in R} \sum_{k \in R} \delta(c,k) \frac{W(u,c,k)}{P(|u|,2)},$$

where $|u|$ is the cardinality (number of judgments[5]) of multiset u, $P(\cdot)$ is the function that computes the number of permutations of the judgments, and $W(u,c,k)$ is the number of ways to make a pair containing c and k from multiset u.

The average disagreement is then computed for each item individually and averaged across all items (weighted by the number of judgments per item) to compute the so-called *observed disagreement*:

$$D_o = \sum_{u \in U} \frac{|u|}{|V|} \mathrm{D}(u),$$

where V is a combined multiset (using the multiset sum $\uplus$) containing the judgments of all items:

$$V = \biguplus_{u \in U} u.$$

The *expected disagreement* is the average disagreement of all possible judgment pairs: $D_e = \mathrm{D}(V)$. Finally, Krippendorff's α can be computed as:

$$\alpha = 1 - \frac{D_o}{D_e}.$$

For example, let us consider the following data, where 4 over 5 assessors agree:

$$X = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

The set of responses R here is $\{0,1\}$. We can choose a function for δ as follow:

$$\delta(c,k) = \begin{cases} c = k & 0 \\ c \neq k & 1. \end{cases}$$

We then collect the results of the observations as a multiset of item multisets:

$$\{\{0,0,0,0,1\},\{1,1,1,1,0\}\},$$

[5]The number of judgments for an item has to be larger than one, otherwise the agreement for that item is undefined.

from which we can determine:

$$V = \{0, 0, 0, 0, 1, 1, 1, 1, 1, 0\}.$$

Now it is possible to compute the expected disagreement:

$$D_e = D(\{0, 0, 0, 0, 1, 1, 1, 1, 1, 0\}),$$

the summations is computed over all possible pairs of values of R, which in this case are (1, 1), (1, 0), (0, 1) and (0, 0).

Consider the term for the pair (1, 1):

$$\delta(1,1)\frac{\mathrm{W(V},1,1)}{\mathrm{P}(|\mathrm{V}|,2)}) = 0,$$

since $\delta(1, 1) = 0$. With the same reasoning we can conclude the contribution from (0, 0) is also 0.

Consider the term corresponding to the pair (1, 0):

$$\delta(1,0)\frac{\mathrm{W(V},1,0)}{\mathrm{P}(|\mathrm{V}|,2)}) = 1\frac{5\times 5}{90} = \frac{5}{18},$$

that is invariant to the order of the pair and thus the contribution due to (0, 1) is also $\frac{5}{18}$.

Now we can compute:

$$D_e = 0 + \frac{5}{18} + \frac{5}{18} + 0 = \frac{5}{9},$$

and the observed disagreement is:

$$\begin{aligned} D_o &= \left(\frac{5}{10}D(\{0,0,0,0,1\}) + \frac{5}{10}D(\{1,1,1,1,0\})\right) \\ &= \left(\frac{5}{10}\left(2\frac{1\times 4}{20}\right) + \frac{5}{10}\left(2\frac{4\times 1}{20}\right)\right) = \frac{2}{5}. \end{aligned}$$

Using these results we can calculate:

$$\alpha = 1 - \frac{2}{5}\cdot\frac{9}{5} = \frac{7}{25} = 0.28.$$

According to (Kripendorff 2004), the data are not reliable, because $\alpha < 0.667$, even if four over five assessors agree for both items.

References

Aroyo, L., and Welty, C. 2014. The three sides of CrowdTruth. *Journal of Human Computation* 1:31–34.

Banerjee, M.; Capozzoli, M.; McSweeney, L.; and Sinha, D. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics* 27(1):3–23.

Bloch, D. A., and Kraemer, H. C. 1989. 2 x 2 kappa coefficients: measures of agreement or association. *Biometrics* 269–287.

Byrt, T.; Bishop, J.; and Carlin, J. B. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology* 46(5):423–429.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213.

Feinstein, A. R., and Cicchetti, D. V. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology* 43(6):543–549.

Fleiss, J. L., and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33(3):613–619.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

Gadiraju, U.; Kawase, R.; and Dietze, S. 2014. A taxonomy of microtasks on the web. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*, 218–223.

Gadiraju, U.; Yang, J.; and Bozzon, A. 2017. Clarity is a worthwhile quality – on the role of task clarity in microtask crowdsourcing. In *28th ACM Conference on Hypertext and Social Media, HT '17, Prague, Czech Republic, July 4-7, 2017*. ACM.

Honour, D. 2016. What is Krippendorff's Alpha. https://github.com/foolswood/krippendorffs_alpha.

Hutchinson, T. 1993. Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. *Research in nursing & health* 16(4):313–316.

Kripendorff, K. 2004. Reliability in content analysis: Some common misconceptions. *Human Communications Research* 30:411–433.

Krippendorff, K. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)* 43.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* 159–174.

O'Connell, D. L., and Dobson, A. J. 1984. General observer-agreement measures on individual subjects and groups of subjects. *Biometrics* 973–983.

Paritosh, P. 2012. Human Computation Must Be Reproducible. In *CrowdSearch*, 20–25.

Salvatier, J.; Wiecki, T. V.; and Fonnesbeck, C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2:e55.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly* 321–325.

Smithson, M., and Verkuilen, J. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods* 11(1):54.

Zhao, X.; Liu, J. S.; and Deng, K. 2013. Assumptions behind intercoder reliability indices. *Annals of the International Communication Association* 36(1):419–480.