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

Kevin Roitero (and Stefano Mizzaro)

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

This publication is based upon work from COST Action CA16105, supported by COST (European Cooperation in Science and Technology).

Lisbon, 15th March 2019

but the real title should have been...

Kevin Roitero (and Stefano Mizzaro)

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

This publication is based upon work from COST Action CA16105, supported by COST (European Cooperation in Science and Technology).

Lisbon, 15th March 2019

The Elephant in the Room

Kevin Roitero (and Stefano Mizzaro)

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

This publication is based upon work from COST Action CA16105, supported by COST (European Cooperation in Science and Technology).

Lisbon, 15th March 2019

This Talk is Based on the Following Paper

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

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Gianluca Demartini and Stefano Mizzaro.

Proceedings of the The fifth AAAI Conference on Human Computation and Crowdsourcing, AAAI HCOMP 2017. Quebec City, Canada. October 24-26 2017.

url: *https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/viewFile/15927/15258*

Setting

- micro-task crowdsourcing
- many workers do the same task
- agreement among workers can / should be leveraged
- leveraging agreement can be useful for:
  - estimating the reliability of collected data
  - understanding behavior of the workers

Agreement Formalization

$$
\begin{array}{cc}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & r_{i,j} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}
$$

Agreement Formalization

$$
\begin{array}{cc}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & r_{i,j} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}
$$

N Items

Agreement Formalization

$$
\begin{array}{cc}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & r_{i,j} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}
$$

N Items

M Assessors

Agreement Formalization

$$
\begin{array}{cc}
 & \begin{matrix} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{matrix} \\
\begin{matrix} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{matrix} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & \boxed{r_{i,j}} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}
$$

N Items

Rating of assessor *i* for item *j*

M Assessors

Agreement Formalization

$$
\begin{array}{c|cccccc}
 & \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \\
\hline
\text{Assessor}_1 & r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
\text{Assessor}_2 & & r_{2,2} & \cdots & & \cdots & r_{2,N} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
\text{Assessor}_i & r_{i,1} & & & \boxed{r_{i,j}} & \cdots & r_{i,N} \\
\vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \\
\text{Assessor}_M & r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{array}
$$

N Items

M Assessors

Rating of assessor *i* for item *j*

This matrix is often **very** sparse in crowdsourcing

There are Several Agreement Measures

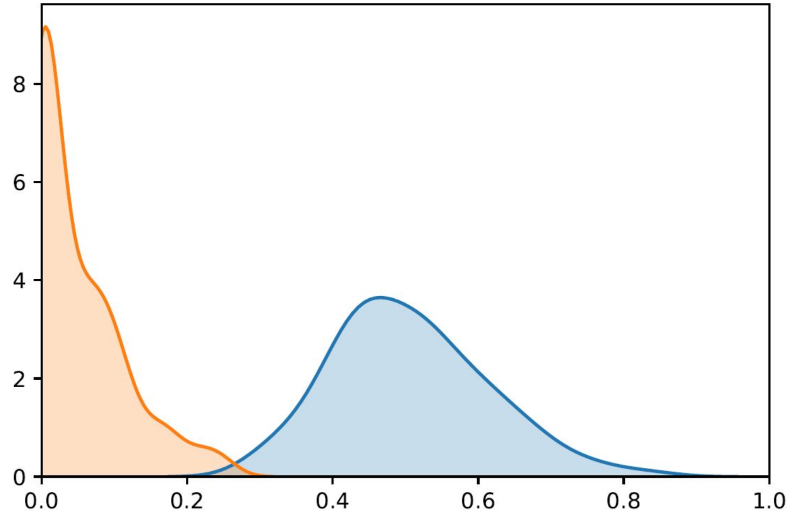
- Percentage Agreement (PA)
- Scott’s π
- Cohen’s κ
- Intraclass Correlation Coefficient (ICC)
- Fleiss κ
- Krippendorff’s Alpha

Current Agreement Measures Are Inadequate

- measures often borrowed from other scenarios with **different assumptions** (which usually do not hold for crowdsourcing):
  - one assessor rates all items
  - all assessors rate all items
  - limited and fixed (= known) number of assessors
- measures are often designed for estimating **data reliability**, not **agreement**
  - **reliability**: the capacity of any measurement tool to differentiate between respondents when measured twice under the same conditions. [Berchtold]
  - **agreement**: the capacity of any other measurement tool applied twice on the same respondents under the same conditions to provide strictly identical results. [Berchtold]
  - reliability can be considered as a necessary but not sufficient condition to demonstrate agreement. [Berchtold]

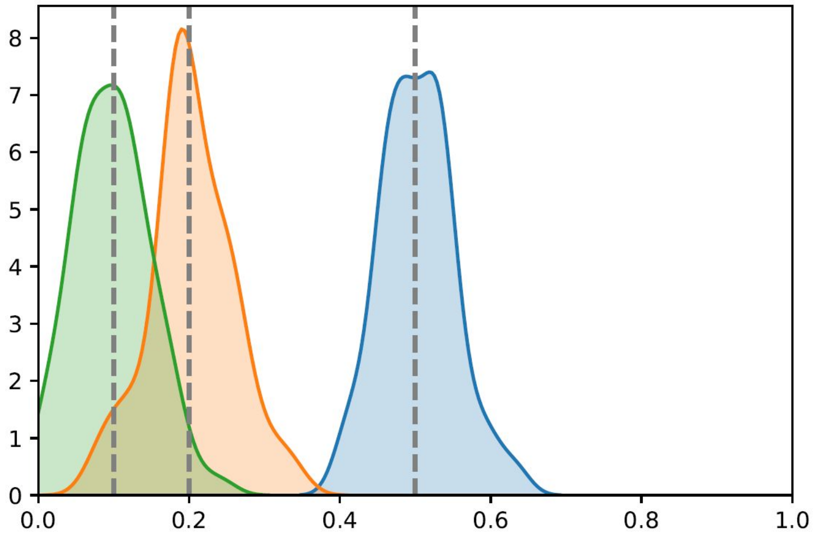
Problems

- there is more variability of judgments in the centre of the scale w.r.t. scale boundaries.
  → can lead to over-estimate agreement close to scale boundaries.



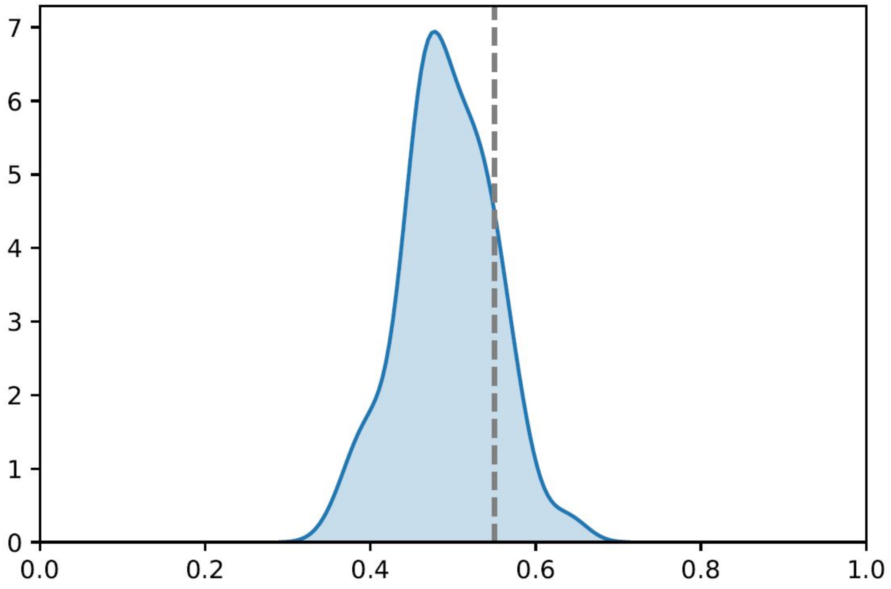
Problems

- the concentration point can be different for different items
  → can lead to over/under-estimate agreement



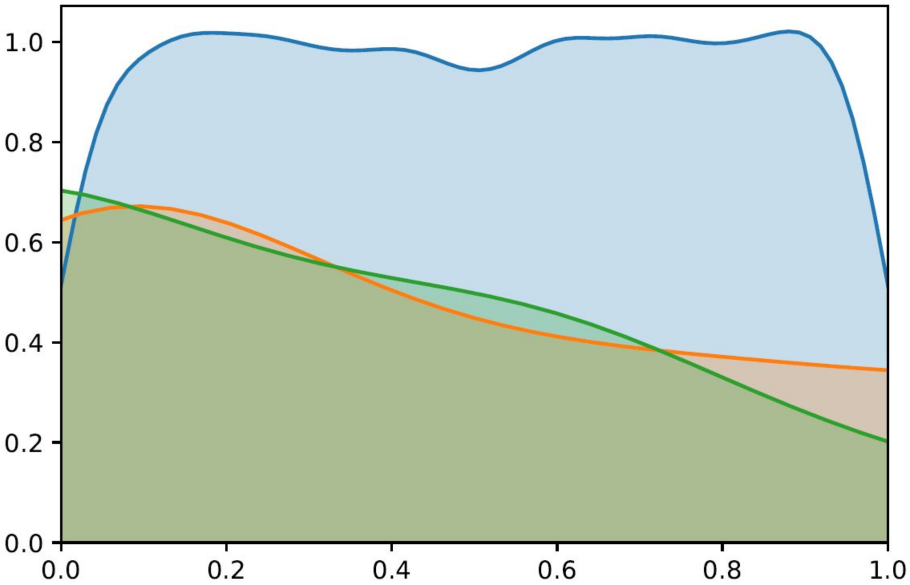
Problems

- additional information is often not considered (e.g., gold questions)



Problems

- different ideas of “**agreement by chance**” definition
- correction by chance assumptions are often violated in crowdsourcing setting



Real Problems with State-of-the-Art Measures

- Percentage Agreement (PA)
  - does not consider agreement by chance
  - works only with nominal data
  - depends on the scale granularity (can not compare different scales)
- Scott's π and Cohen's κ
  - work only with two assessors
  - work only with nominal data
- Intraclass Correlation Coefficient (ICC)
  - assessor have same marginal probability of an answer (not true in crowdsourcing)
  - equivalent to weighted Cohen's κ
- Fleiss κ
  - Generalizes κ to multiple assessors (i.e., shares the same issues)

Real Problems with State-of-the-Art Measures

- Krippendorff's Alpha: an attempt to generalize previous metrics
    - **Random guessing can have high agreement**
    - **Random guessing may have more agreement than honest coding**
    - High agreement, low reliability
    - Zero change in percentage agreement causing radical drop in reliability.
    - **Eliminating disagreements does not improve agreement**
    - Honest work as bad as coin flipping.
    - Two datasets: same quality, same agreement; but higher reliability in one.
    - punishing larger sample and replicability (i.e., data quantity dependent)
    - **"reverse answer" problem** $([1,0,0,0,1] \neq [0,1,1,1,0])$

(a complete overview and all the mathematical details are available in our paper)

Our Measure: Φ

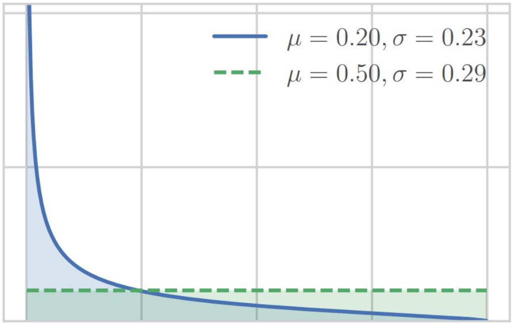
- *agreement* definition as a key point:

  "agreement is the amount of concentration around a data value"
- if we do not observe agreement (i.e., concentration around a point), we have disagreement, treated as negative agreement in our measure
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:

Our Measure: Φ

- *agreement* definition as a key point:

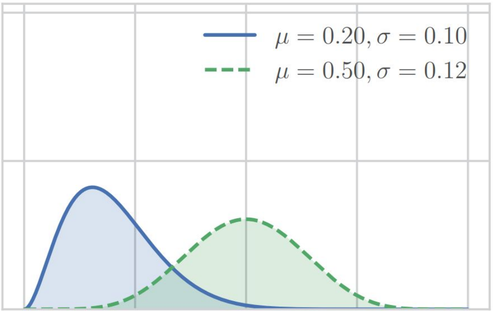
  “agreement is the amount of concentration around a data value”
- if we do not observe agreement (i.e., concentration around a point), we have disagreement, treated as negative agreement in our measure
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:
  - random judgments → flat distribution



Our Measure: Φ

- *agreement* definition as a key point:

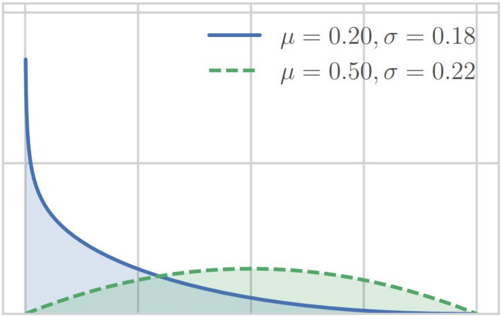
  "agreement is the amount of concentration around a data value"
- if we do not observe agreement (i.e., concentration around a point), we have disagreement, treated as negative agreement in our measure
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:
  - random judgments → flat distribution
  - agreement → bell-shaped distribution



Our Measure: Φ

- *agreement* definition as a key point:

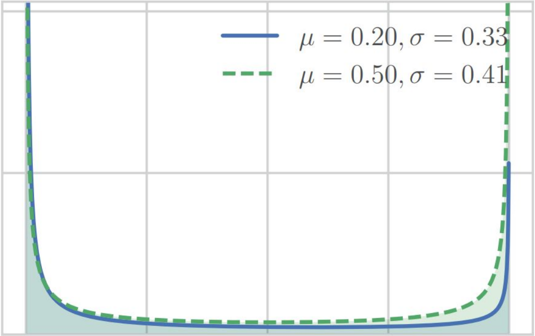
  "agreement is the amount of concentration around a data value"
- if we do not observe agreement (i.e., concentration around a point), we have disagreement, treated as negative agreement in our measure
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:
  - random judgments → flat distribution
  - agreement → bell-shaped distribution
  - agreement around scale boundaries → J-distribution



Our Measure: Φ

- *agreement* definition as a key point:

  "agreement is the amount of concentration around a data value"
- if we do not observe agreement (i.e., concentration around a point), we have disagreement, treated as negative agreement in our measure
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:
  - random judgments → flat distribution
  - agreement → bell-shaped distribution
  - agreement around scale boundaries → J-distribution
  - disagreement → U shaped distribution



Our Measure: Φ

- *agreement* definition as a key point:

  "agreement is the amount of concentration around a data value"
- if we do not observe agreement (i.e., concentration around a point), we have disagreement, treated as negative agreement in our measure
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:
  - random judgments → flat distribution
  - agreement → bell-shaped distribution
  - agreement around scale boundaries → J-distribution
  - disagreement → U shaped distribution
- we should have a minimal number of parameters, to avoid overfitting

Our Measure: Φ

- we use a Beta distribution to model our scenario: $B(a, b)$
- we re-parametrize the distribution in terms of the mean value μ and the precision p as $\mu = \frac{a}{a+b}; p = a + b$
- now, we can treat separately mean and dispersion
- we can have a metric that is agnostic of the mean value
- then, we transform to have values in the [-1, +1] range:

$$\Phi = 1 - 2^{\frac{-p\, log2}{2}}$$

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

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

probability of observing the mean values, with a common dispersion, given the observed data

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

probability of observing the mean values, with a common dispersion, given the observed data

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

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

probability of observing the mean values, with a common dispersion, given the observed data

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

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

probability of observing the mean values, with a common dispersion, given the observed data

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

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

probability of observing the mean values, with a common dispersion, given the observed data

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

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

probability of observing the mean values, with a common dispersion, given the observed data

- Then, we estimate Φ using

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

Our Measure: Φ

- we use Bayesian inference to compute Φ:

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

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

probability of observing the mean values, with a common dispersion, given the observed data

- Then, we estimate Φ using

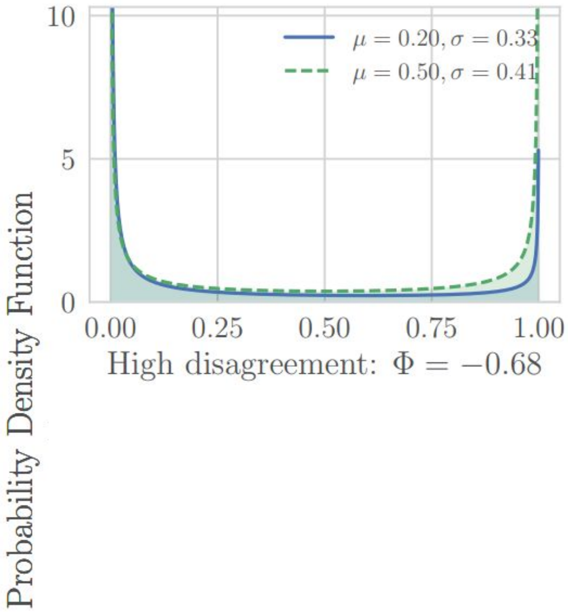
$$\hat{\Phi} = \arg\max_{\Phi} \mathrm{P}(\vec{\mu}, \Phi | X).$$

the formula can change to incorporate custom ground truth

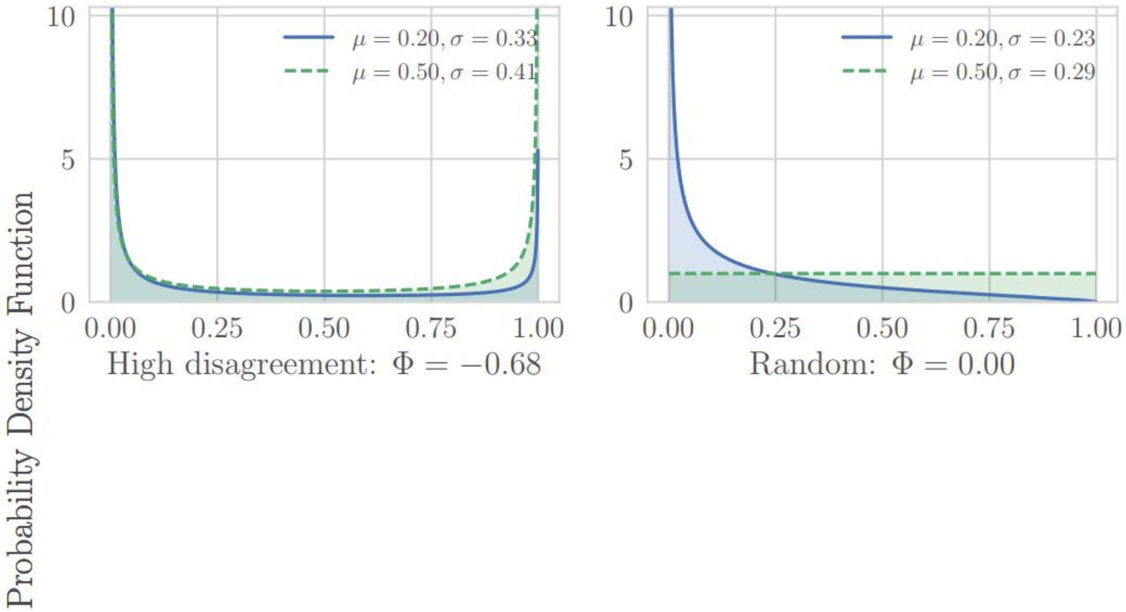
Φ Interpretation

- **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

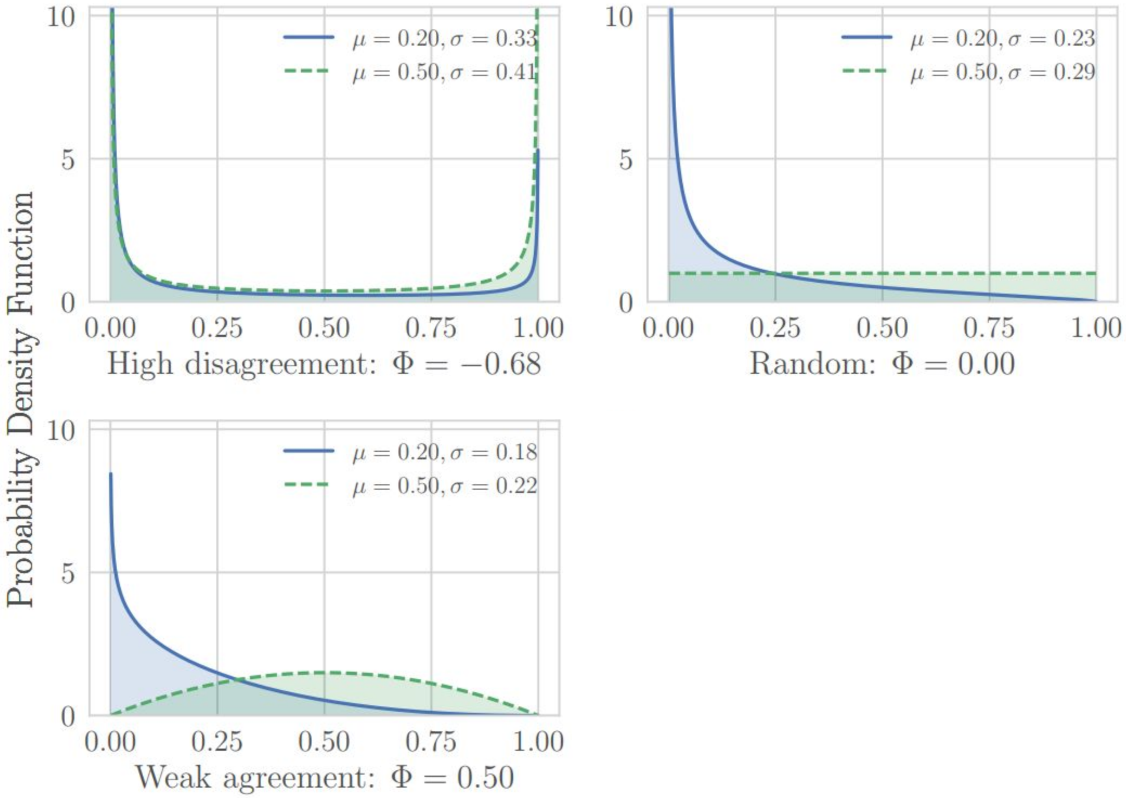
Examples of Φ Shapes



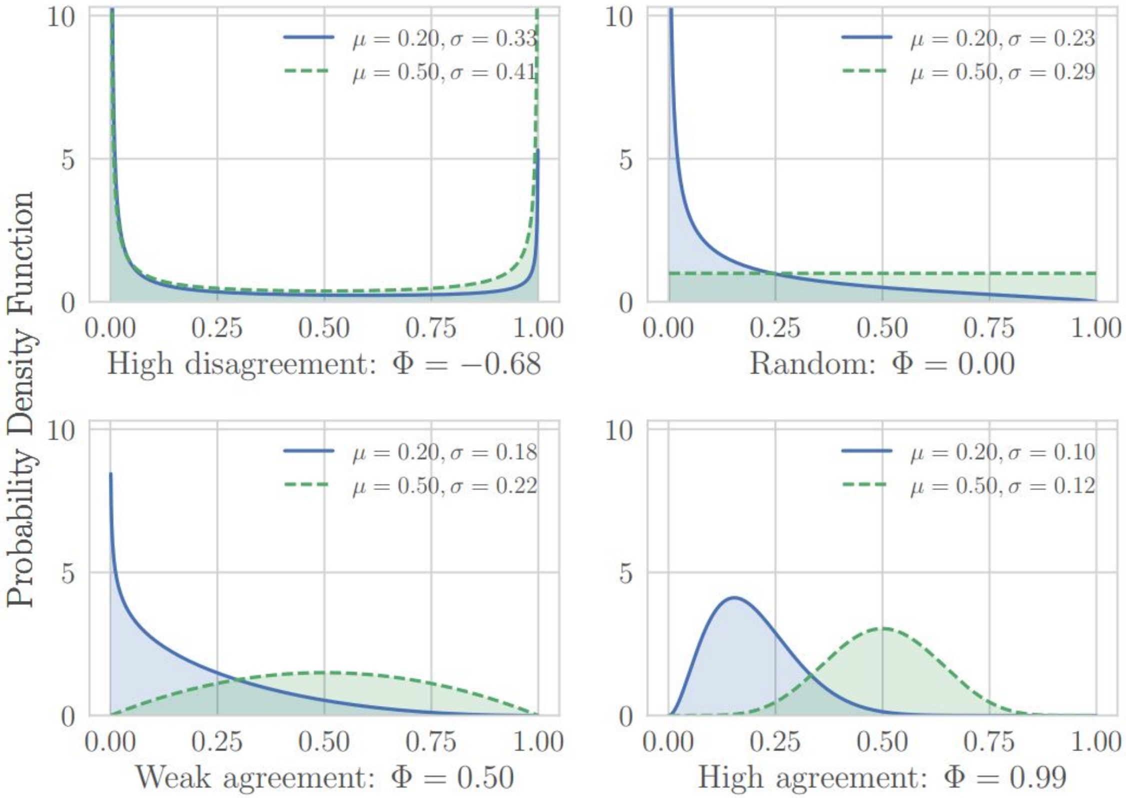
Examples of Φ Shapes



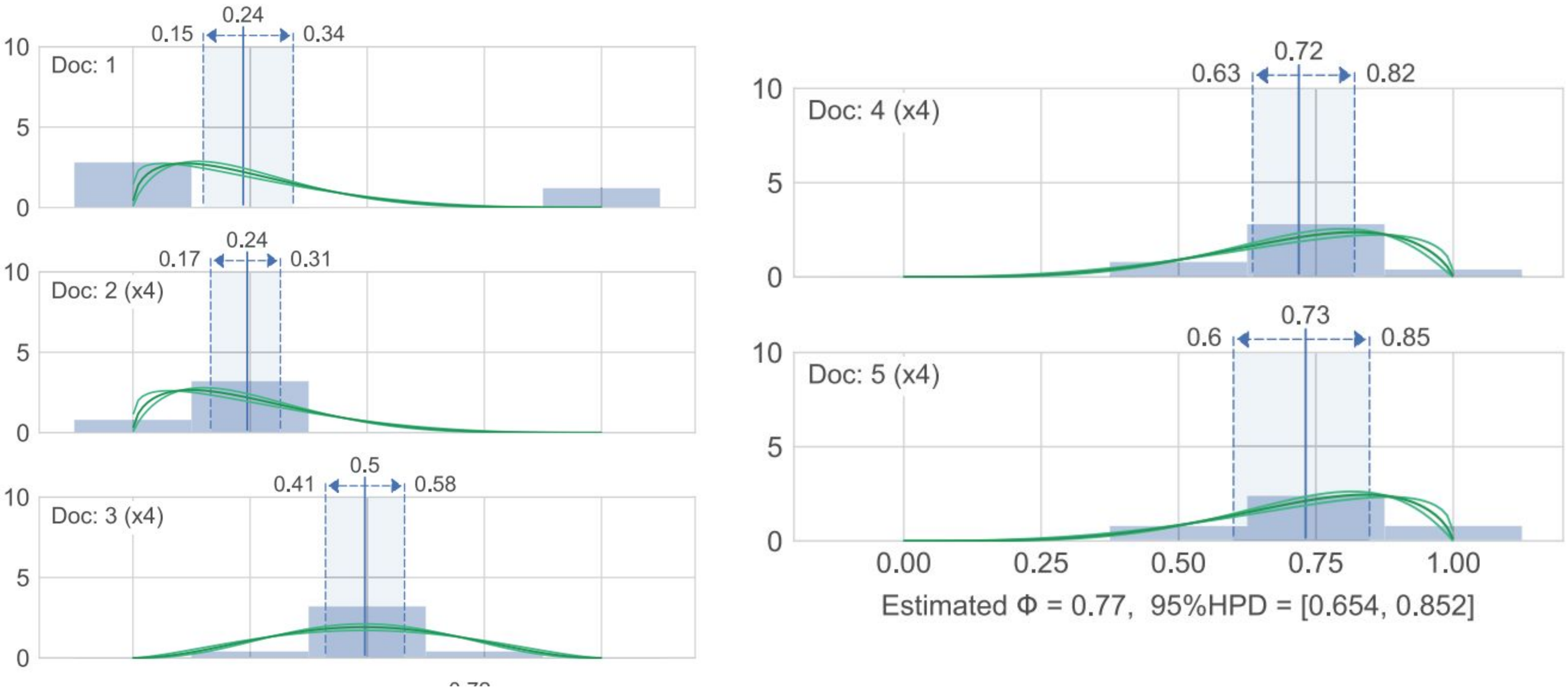
Examples of Φ Shapes



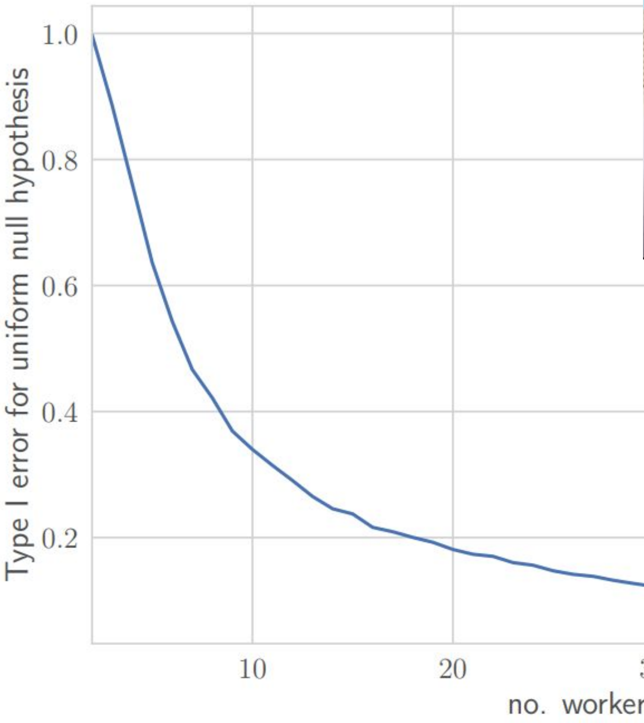
Examples of Φ Shapes

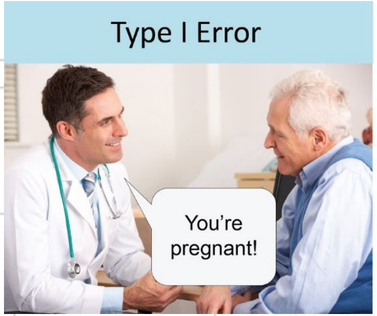


Φ in Action on Real Data

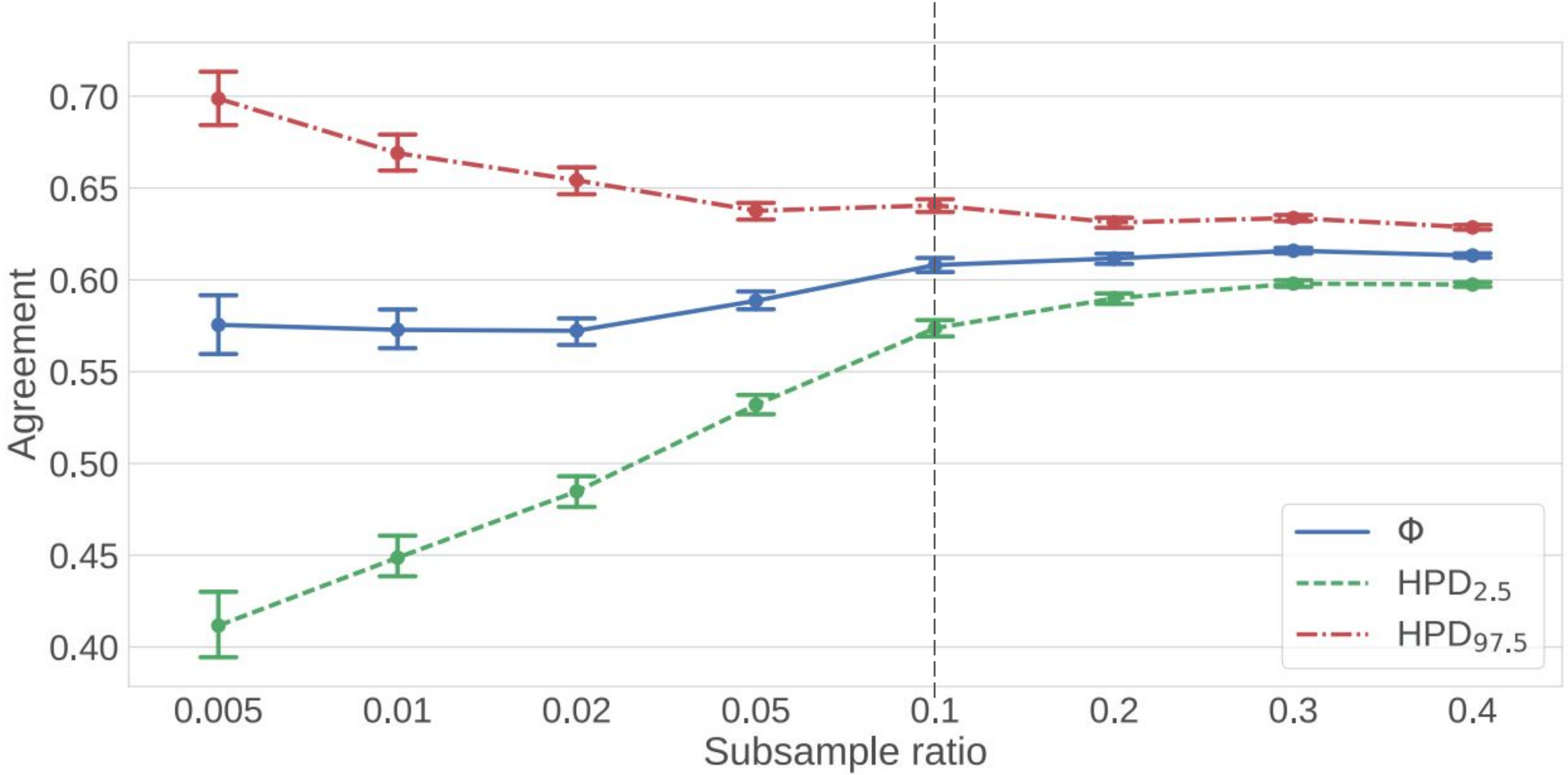


Robustness of Φ





Confidence Interval, Robustness of Φ



Done / Ongoing / Future Developments

- Incorporate agreement in metrics used for evaluation
- Incorporate agreement into aggregation methods
- Extend / fine tune Φ for different scales (categorical, ratio, etc.)
- Deal with bias / reputation: different weights for different items / assessors

Take Home Messages

- Φ is a new agreement measure
- Φ has a set of nice properties that makes it suitable for different (crowdsourcing) scenarios
- Φ can be customized and adapted to different situations

Φ Properties Summary

- We have a **confidence interval** for the measure
- If we have **prior knowledge** on the domain (e.g., gold question), we could use that in the computation of the metric
(by adding a set of priors to the model)
- We can deal with items having **different concentration points**

Resources

- Paper: https://www.aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/viewFile/15927/15258
- “Follow the Crowd” article: https://blog.humancomputation.com/?p=9756
- Python-library (pip) and GitHub Repository: https://pypi.org/project/agreement-phi/
- Live Demo / Online Tool: http://agreement-measure.sheffield.ac.uk/