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

Kevin Roitero (and Stefano Mizzaro)
but the real title should have been...

Kevin Roitero (and Stefano Mizzaro)
The Elephant in the Room

Kevin Roitero (and Stefano Mizzaro)

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

Lisbon, 15th March 2019
Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing.
Alessandro Checco, Kevin Roitero, Eddy Maddalena, Gianluca Demartini and Stefano Mizzaro.
Quebec City, Canada. October 24-26 2017.

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

<table>
<thead>
<tr>
<th></th>
<th>Item$_{1}$</th>
<th>Item$_{2}$</th>
<th>\ldots</th>
<th>Item$_{j}$</th>
<th>\ldots</th>
<th>Item$_{N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor$_{1}$</td>
<td>$r_{1,1}$</td>
<td>$r_{1,2}$</td>
<td>\ldots</td>
<td>$r_{1,j}$</td>
<td>\ldots</td>
<td>$r_{1,N}$</td>
</tr>
<tr>
<td>Assessor$_{2}$</td>
<td>$r_{2,1}$</td>
<td>$r_{2,2}$</td>
<td>\ldots</td>
<td>$r_{2,j}$</td>
<td>\ldots</td>
<td>$r_{2,N}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Assessor$_{i}$</td>
<td>$r_{i,1}$</td>
<td>$r_{i,2}$</td>
<td>\ldots</td>
<td>$r_{i,j}$</td>
<td>\ldots</td>
<td>$r_{i,N}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Assessor$_{M}$</td>
<td>$r_{M,1}$</td>
<td>$r_{M,2}$</td>
<td>\ldots</td>
<td>$r_{M,j}$</td>
<td>\ldots</td>
<td>$r_{M,N}$</td>
</tr>
</tbody>
</table>
## Agreement Formalization

<table>
<thead>
<tr>
<th>Assessor(i)</th>
<th>Item(_1)</th>
<th>Item(_2)</th>
<th>(\cdots)</th>
<th>Item(_j)</th>
<th>(\cdots)</th>
<th>Item(_N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor(_1)</td>
<td>(r_{1,1})</td>
<td>(r_{1,2})</td>
<td>(\cdots)</td>
<td>(r_{1,j})</td>
<td>(\cdots)</td>
<td>(r_{1,N})</td>
</tr>
<tr>
<td>Assessor(_2)</td>
<td>(r_{2,1})</td>
<td>(r_{2,2})</td>
<td>(\cdots)</td>
<td>(r_{2,j})</td>
<td>(\cdots)</td>
<td>(r_{2,N})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\cdots)</td>
<td>(\vdots)</td>
<td>(\cdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>Assessor(_i)</td>
<td>(r_{i,1})</td>
<td>(r_{i,2})</td>
<td>(\cdots)</td>
<td>(r_{i,j})</td>
<td>(\cdots)</td>
<td>(r_{i,N})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\cdots)</td>
<td>(\vdots)</td>
<td>(\cdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>Assessor(_M)</td>
<td>(r_{M,1})</td>
<td>(r_{M,2})</td>
<td>(\cdots)</td>
<td>(r_{M,j})</td>
<td>(\cdots)</td>
<td>(r_{M,N})</td>
</tr>
</tbody>
</table>

\(N\) Items
# Agreement Formalization

A table illustrating the agreement of $M$ assessors with $N$ items, where $r_{ij}$ represents the rating of the $i$th assessor for the $j$th item.

<table>
<thead>
<tr>
<th></th>
<th>Item 1</th>
<th>Item 2</th>
<th>...</th>
<th>Item $j$</th>
<th>...</th>
<th>Item $N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor 1</td>
<td>$r_{1,1}$</td>
<td>$r_{1,2}$</td>
<td>...</td>
<td>$r_{1,j}$</td>
<td>...</td>
<td>$r_{1,N}$</td>
</tr>
<tr>
<td>Assessor 2</td>
<td>$r_{2,1}$</td>
<td>$r_{2,2}$</td>
<td>...</td>
<td>$r_{2,j}$</td>
<td>...</td>
<td>$r_{2,N}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Assessor $i$</td>
<td>$r_{i,1}$</td>
<td>$r_{i,2}$</td>
<td>...</td>
<td>$r_{i,j}$</td>
<td>...</td>
<td>$r_{i,N}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Assessor $M$</td>
<td>$r_{M,1}$</td>
<td>$r_{M,2}$</td>
<td>...</td>
<td>$r_{M,j}$</td>
<td>...</td>
<td>$r_{M,N}$</td>
</tr>
</tbody>
</table>

Each assessor provides a rating for each item, allowing for the formalization of agreement across multiple assessors and items.
Agreement Formalization

<table>
<thead>
<tr>
<th>Assessor $i$</th>
<th>$r_{i,1}$</th>
<th>$r_{i,2}$</th>
<th>$\cdots$</th>
<th>$r_{i,j}$</th>
<th>$\cdots$</th>
<th>$r_{i,N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor $1$</td>
<td>$r_{1,1}$</td>
<td>$r_{1,2}$</td>
<td>$\cdots$</td>
<td>$r_{1,j}$</td>
<td>$\cdots$</td>
<td>$r_{1,N}$</td>
</tr>
<tr>
<td>Assessor $2$</td>
<td>$r_{2,1}$</td>
<td>$r_{2,2}$</td>
<td>$\cdots$</td>
<td>$r_{2,j}$</td>
<td>$\cdots$</td>
<td>$r_{2,N}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>Assessor $M$</td>
<td>$r_{M,1}$</td>
<td>$r_{M,2}$</td>
<td>$\cdots$</td>
<td>$r_{M,j}$</td>
<td>$\cdots$</td>
<td>$r_{M,N}$</td>
</tr>
</tbody>
</table>

- **$N$ Items**
- **$M$ Assessors**
- **Rating of assessor $i$ for item $j$**
## Agreement Formalization

This matrix is often **very** sparse in crowdsourcing.
There are Several Agreement Measures

- Percentage Agreement (PA)
- Scott’s $\pi$
- Cohen’s $\kappa$
- Intraclass Correlation Coefficient (ICC)
- Fleiss $\kappa$
- 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 $\pi$ and Cohen’s $\kappa$**
  - 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 $\kappa$

- **Fleiss $\kappa$**
  - Generalizes $\kappa$ 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: $\Phi$

- *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: $\Phi$

- *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: $\Phi$

- *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 $\rightarrow$ flat distribution
  - agreement $\rightarrow$ bell-shaped distribution

![Graph showing distribution shapes](image)
Our Measure: $\Phi$

- \textit{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 $\rightarrow$ flat distribution
  - agreement $\rightarrow$ bell-shaped distribution
  - agreement around scale boundaries $\rightarrow$ J-distribution

![Graph showing distribution fitting and dispersion](image)
Our Measure: \( \Phi \)

- *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 \( \rightarrow \) flat distribution
  - agreement \( \rightarrow \) bell-shaped distribution
  - agreement around scale boundaries \( \rightarrow \) J-distribution
  - disagreement \( \rightarrow \) U shaped distribution
Our Measure: $\Phi$

- **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 $\rightarrow$ flat distribution
  - agreement $\rightarrow$ bell-shaped distribution
  - agreement around scale boundaries $\rightarrow$ J-distribution
  - disagreement $\rightarrow$ U shaped distribution

- we should have a minimal number of parameters, to avoid overfitting
Our Measure: $\Phi$

- we use a Beta distribution to model our scenario: $B(a, b)$
- we re-parametrize the distribution in terms of the mean value $\mu$ 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:

$$\Phi = 1 - 2^{-p \log 2}$$
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

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

probability of observing the mean values, with a common dispersion, given the observed data
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

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

probability of observing the mean values, with a common dispersion, given the observed data.
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

probability of observing the mean values, with a common dispersion, given the observed data
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

probability of observing the mean values, with a common dispersion, given the observed data
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

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

probability of observing the mean values, with a common dispersion, given the observed data
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

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

- Then, we estimate $\Phi$ using

$$\hat{\Phi} = \arg \max_{\Phi} P(\mu, \Phi | X).$$
Our Measure: $\Phi$

- we use Bayesian inference to compute $\Phi$:

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

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

- Then, we estimate $\Phi$ using

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

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

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 $\Phi$ grows
Examples of $\Phi$ Shapes
Examples of $\Phi$ Shapes
Examples of \( \Phi \) Shapes

- High disagreement: \( \Phi = -0.68 \)
- Random: \( \Phi = 0.00 \)
- Weak agreement: \( \Phi = 0.50 \)
Examples of $\Phi$ Shapes

- High disagreement: $\Phi = -0.68$
- Random: $\Phi = 0.00$
- Weak agreement: $\Phi = 0.50$
- High agreement: $\Phi = 0.99$
Φ in Action on Real Data

Estimated Φ = 0.77, 95% HPD = [0.654, 0.852]
Robustness of $\Phi$
Confidence Interval, Robustness of $\Phi$
Done / Ongoing / Future Developments

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

Take Home Messages

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

- “Follow the Crowd” article: [https://blog.humancomputation.com/?p=9756](https://blog.humancomputation.com/?p=9756)
- Python-library (pip) and GitHub Repository: [https://pypi.org/project/agreement-phi/](https://pypi.org/project/agreement-phi/)
- Live Demo / Online Tool: [http://agreement-measure.sheffield.ac.uk/](http://agreement-measure.sheffield.ac.uk/)