Quality Control
part 2
Different Mechanisms for Quality Control

Aggregation and redundancy
Embedded gold standard data
Economic incentives
Reputation systems
Statistical models
Reputation systems

Mechanical Turk uses a reputation system
Each Turker has a small number of variables associated with them, that are exposed to Requesters
Past approval rate
Number of HITs approved
Has masters qualification (photo moderation/ categorization master)
## Pros and Cons of MTurk’s reputation system

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gives a bit information about what other Requesters thought of a Worker</td>
<td>Reasons for rejections not shared; Weights all Requesters equally</td>
</tr>
<tr>
<td>Allows you to select Amazon’s master’s qualification, which is given to experienced Workers</td>
<td>It is not clear who gets the master’s qual. No way to share other qualifications.</td>
</tr>
<tr>
<td></td>
<td>Asymmetric: applies only to Workers, with no way to rate Requesters</td>
</tr>
</tbody>
</table>
Confederated Trust

Acceptance rate doesn’t show how good a worker is at a particular task. Qualifications like the “photo moderation master’s” may show this. However, there is no way to share this information with other requester
Lots of reinventing the wheel
Confederated Trust

Do you think it would be useful to share qualifications among requesters? How would you do it?
Asymmetric reputation systems

No way for Turkers to rate requesters, and see beforehand who is scrupulous

Turkers have built their own external tools for this like TurkOpticon

No way to see whether a Turkers high rating comes from good Requesters
The Public Group
A1ITPVFB965TV

The Public
Group
A1ITPVFB965TV

The Public
Group
A1ITPVFB965TV

The Public
Group
A1ITPVFB965TV

The Public
Group
A1ITPVFB965TV

Took a leap of faith on this requester and was rewarded with a 50% reject rate and a broken search feature and no feedback. Would not recommend, even if you have thousands of HITs under your belt to cushion the inevitable rejections.

Aug 29 2013 | KBH19 | flag | comment

Arbitrarily rejected over half of the hits I submitted, and then banned me from submitting any more hits for them. I suppose that's a blessing in disguise though, as I had no intention of doing any for them again after the first batch of rejections.

Aug 21 2013 | bour...@g... | flag | comment

Their HIT is very unclear. There is an option to browse for the result, but it does not work.

Aug 20 2013 | jeff...@g... | flag | comment

Scores based on 81 reviews
Terms of Service violation flags: 0
Report your experience with this requester »

Communication: 1.17 / 5
Generosity: 1.73 / 5
Fairness: 1.39 / 5
Promptness: 1.86 / 5

What do these scores mean?

©2005-2013 Amazon.com, Inc. or its Affiliates
# Qualitative v Quantitative

<table>
<thead>
<tr>
<th>TurkOpticon's qualitative attributes</th>
<th>CrowdWorker's quantitative equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>promptness</strong>: How promptly has this requester approved your work and paid?</td>
<td><strong>Expected time to payment</strong>: On average, how much time elapses between submitting work to this Requester and receiving payment?</td>
</tr>
<tr>
<td><strong>generosity</strong>: How well has this requester paid for the amount of time their HITs take?</td>
<td><strong>Average hourly rate</strong>: What is the average hourly rate that other Turkers make when they do this requester's HITs?</td>
</tr>
<tr>
<td><strong>fairness</strong>: How fair has this requester been in approving or rejecting your work?</td>
<td><strong>Approval/rejection rates</strong>: What percent of assignments does this Requester approve? What percent of first-time Workers get any work rejected?</td>
</tr>
<tr>
<td><strong>communicativity</strong>: How responsive has this requester been to communications or concerns you have raised?</td>
<td><strong>Reasons for rejection</strong>: Archive of all of the reasons for Workers being rejected or blocked by this Requester.</td>
</tr>
</tbody>
</table>
Amazon’s other reputation system

Amazon has another reputation system in place for its online stores. Amazon allows anyone to list and sell items through its site, and to set their own prices. These can be individuals selling used goods, or independent 3rd party sellers who use Amazon to reach a larger customer base. How does Amazon ensure good customer experience?
Feedback from buyers

How satisfied were you with how your order was packaged and shipped?
If you contacted the third-party seller, did you get good customer service and prompt resolution?
Would you buy from this third-party seller again?
Westinghouse Lighting 7214100
Harmony Two-Light 48-Inch Two-Blade Indoor Ceiling Fan, Brushed Nickel with Opal Frosted Glass
by Westinghouse

4.5 stars (42 customer reviews) | 11 answered questions
<table>
<thead>
<tr>
<th>Price + Shipping</th>
<th>Condition</th>
<th>Seller Information</th>
<th>Buying Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>$128.69</td>
<td>New</td>
<td><a href="https://www.amazon.com">amazon.com</a></td>
<td>Add to cart or Turn on 1-Click</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In Stock.</td>
<td>to use your Amazon Prime benefits.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Free Two-day Shipping: Get it Wednesday, October 2 (order within)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Domestic shipping rates and return policy.</td>
<td></td>
</tr>
<tr>
<td>$128.69</td>
<td>New</td>
<td><a href="https://www.plumbersurplus.com">PlumberSurplus</a></td>
<td>Add to cart or Sign in to turn on 1-Click</td>
</tr>
<tr>
<td>+ $24.32 shipping</td>
<td></td>
<td>★★★★★: 91% positive over the past 12 months. (12,817 total ratings)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ships in 1-2 business days. Expedited shipping available. Domestic shipping rates and return policy.</td>
<td></td>
</tr>
<tr>
<td>$128.69</td>
<td>New</td>
<td><a href="https://www.remodelr.com">Remodelr</a></td>
<td>Add to cart or Sign in to turn on 1-Click</td>
</tr>
<tr>
<td>+ $24.32 shipping</td>
<td></td>
<td>★★★★★: 90% positive over the past 12 months. (1,855 total ratings)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ships in 1-2 business days. Domestic shipping rates and return policy.</td>
<td></td>
</tr>
<tr>
<td>$148.99</td>
<td>New</td>
<td><a href="https://www.elastoflex.com">Elasto Flex</a></td>
<td>Add to cart or Sign in to turn on 1-Click</td>
</tr>
<tr>
<td>+ $24.19 shipping</td>
<td></td>
<td>★★★★★: 97% positive over the past 12 months. (163,508 total ratings)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Usually ships within 3 - 4 business days. Domestic shipping rates and return policy.</td>
<td></td>
</tr>
<tr>
<td>$202.90</td>
<td>New</td>
<td><a href="https://www.delmarfans.com">DEL MAR</a></td>
<td>Add to cart or Sign in to turn on 1-Click</td>
</tr>
<tr>
<td>FREE Shipping</td>
<td></td>
<td>★★★★★: 98% positive over the past 12 months. (7,007 total ratings)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Usually ships within 4 - 5 business days. Domestic shipping rates and return policy.</td>
<td></td>
</tr>
</tbody>
</table>

---

Penn Engineering
Recent Feedback: ★★★★★
4.6 stars over the past 12 months (573 ratings)

5/5: "good transaction, love the lock"
Sophia D., September 22, 2013

5/5: "Awesome experience."
Yadira Morejon, September 22, 2013

5/5: "Exactly what I needed, especially the color matched perfectly. Thank you."
mufasa, September 20, 2013

5/5: "Item was as described"
Thomas F., September 20, 2013

5/5: "Great seller, great item! Fast service too!!"
DB, September 20, 2013

2/5: "arrived bent"
ATD, September 20, 2013

Seller Response: We were not aware of any issue involving this customer's order. We have reached out to the customer to see how we may assist them in a return for a full refund, or a replacement of the damaged item.
Date: September 23, 2013

5/5: "Shipping was a little pricey."

Kylie S., September 19, 2013

5/5: "Good price, high quality."
Joanna Wang, September 18, 2013

5/5: "item was as described seller prompt with service"
Thomas Howell, September 16, 2013

5/5: "works great"
Spencer, September 16, 2013

5/5: "just as described"
Theresa M., September 15, 2013

5/5: "Item was as described, value priced and works great."
Amanda S., September 14, 2013

1/5: "When a seller charges $29.95 in shipping for a package weighing .2 lbs, they are gouging. The sponges cost less than $18. I assume their profit is from the shipping. I could get this shipped for less than $6.00! Never again"
Lana L Miller, September 14, 2013

Seller Response: We apologize the customer is not satisfied with the shipping charges. We have reached out to the customer and offered a discount to them as a one time courtesy.
Date: September 17, 2013
What are the economic implications of poor feedback?
Price premium

Multiple sellers all selling the same item, but at different prices

Price premium is the difference between a cheaper listing and a more expensive listing

When someone opts for the more expensive item, even though it is identical, what is the reason for paying the premium?
Data-driven analysis

Panos Ipeirotis harvested data from Amazon’s website

Gathered **transaction data** by repeatedly visiting listings (every 8 hours) and tracking when one item sold

Gathered **reputation data** for each merchant. Complete history of numerical scores and text-based feedback
Data-driven analysis

Data set gathered over half a year period
Transaction data contains 1,078 merchants, 9,484 unique transactions and 107,922 price premiums
Reputation data contains an average of 4,932 postings for each merchant
NLP + Economics

Quantify the economics impact of sentiment of the feedback evaluations
Using natural language processing techniques to derive semantic orientation
and strength of comments
Method

Each merchant’s reputation is represented using a vector of n-dimensions $X = (X_1, X_2, ... X_N)$

Dimensions were 150 nouns and verbs, values of dimensions could be one of 140 modifiers

$X_1$ is “delivery,” $X_2$ is “packaging,” $X_3$ is “service.”

Feedback 1 “I was impressed by the speedy delivery! Great service!”: (speedy; NULL; great)

Feedback 2 “The item arrived in awful packaging, and the delivery was slow”: (slow; awful; NULL)
Method

Construct a matrix out of all of the feedback for a seller
Weight the more recent feedback more heavily
Calculate how the values of each dimension effect the price premium
Use least-squares regression with fixed effects to predict the price premium
### Highest scoring phrases

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>wonderful experience</td>
<td>$5.86</td>
</tr>
<tr>
<td>outstanding seller</td>
<td>$5.76</td>
</tr>
<tr>
<td>excellant service</td>
<td>$5.27</td>
</tr>
<tr>
<td>lightning delivery</td>
<td>$4.84</td>
</tr>
<tr>
<td>highly recommended</td>
<td>$4.15</td>
</tr>
<tr>
<td>best seller</td>
<td>$3.80</td>
</tr>
<tr>
<td>perfectly packaged</td>
<td>$3.74</td>
</tr>
<tr>
<td>excellent condition</td>
<td>$3.53</td>
</tr>
<tr>
<td>excellent purchase</td>
<td>$3.22</td>
</tr>
<tr>
<td>excellent seller</td>
<td>$2.70</td>
</tr>
<tr>
<td>excellent communication</td>
<td>$2.38</td>
</tr>
<tr>
<td>perfect item</td>
<td>$1.92</td>
</tr>
<tr>
<td>terrific condition</td>
<td>$1.87</td>
</tr>
<tr>
<td>top quality</td>
<td>$1.67</td>
</tr>
<tr>
<td>awesome service</td>
<td>$1.05</td>
</tr>
<tr>
<td>A+++ seller</td>
<td>$1.03</td>
</tr>
<tr>
<td>great merchant</td>
<td>$0.93</td>
</tr>
<tr>
<td>friendly service</td>
<td>$0.81</td>
</tr>
<tr>
<td>easy service</td>
<td>$0.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>never received</td>
<td>-$7.56</td>
</tr>
<tr>
<td>defective product</td>
<td>-$6.82</td>
</tr>
<tr>
<td>horrible experience</td>
<td>-$6.79</td>
</tr>
<tr>
<td>never sent</td>
<td>-$6.69</td>
</tr>
<tr>
<td>never received</td>
<td>-$5.29</td>
</tr>
<tr>
<td>bad experience</td>
<td>-$5.26</td>
</tr>
<tr>
<td>cancelled order</td>
<td>-$5.01</td>
</tr>
<tr>
<td>never responded</td>
<td>-$4.87</td>
</tr>
<tr>
<td>wrong product</td>
<td>-$4.39</td>
</tr>
<tr>
<td>not as advertised</td>
<td>-$3.93</td>
</tr>
<tr>
<td>poor packaging</td>
<td>-$2.92</td>
</tr>
<tr>
<td>late shipping</td>
<td>-$2.89</td>
</tr>
<tr>
<td>wrong item</td>
<td>-$2.50</td>
</tr>
<tr>
<td>not yet received</td>
<td>-$2.35</td>
</tr>
<tr>
<td>still waiting</td>
<td>-$2.25</td>
</tr>
<tr>
<td>wrong address</td>
<td>-$1.54</td>
</tr>
<tr>
<td>never buy</td>
<td>-$1.48</td>
</tr>
</tbody>
</table>
Predicting the merchant who makes the sale

Accuracy

- Price: 55
- Price + Stars: 74
- Price + Stars + Text: 89
- Price + Text: 87
Challenges for Reputation Systems

Not enough people participate
Feedback tends to be overwhelmingly positive
Reports can be dishonest
Reputation systems are undermined if people can change identities easily
People can milk a good reputation
Insufficient participation

Giving feedback for a reputation system contributes to the public good

However, after some information is available it is easy for people to be "free riders" without contributing anything

Early raters take on a transaction cost (Yelpers risk going to bad restaurants with no reviews)

Solutions?
Overwhelmingly positive feedback

99% of all feedback on eBay is positive

Part of the problem is reciprocity

Sellers and buyers evaluate each other

Positive ratings are given in the hopes of getting positive ratings in return

Negative ratings are avoided for fear of getting negative feedback as retaliation
Dishonest reports

**Ballot stuffing** - a seller colludes with buyers to give unfairly high ratings

**Bad mouthing** - collusion to give negative feedback about competitors that they want to drive out of the market
Identity changes

Cheap pseudonyms - easy to disappear and re-register under a new identity with almost zero cost
Can misbehave without paying consequences toward reputation
Value imbalance exploitations

People who want to commit fraud could first invest in building a good reputation.

Ebay exploit: "Riddle for 1¢. No shipping. Positive feedback"

Sellers would take a 29¢ loss to build up positive reputation quickly.
Challenges for Crowdsourcing Markets

Reciprocal systems are worse than 1-sided systems in e-commerce.

In e-commerce, only the sellers are likely to behave opportunistically. No need for reciprocal evaluation.

In crowdsourcing, both sides can be fraudulent. So reciprocal markets are important, but they are hard to get right!
Challenges for Crowdsourcing Markets

In e-commerce markets, it is straightforward for buyers to evaluate the quality of the product when they receive it.

In crowdsourcing markets, verifying the correct answer is sometimes as costly as producing it.

This has the potential to significantly reduce participation and/or accuracy of reviews.
Challenges for Crowdsourcing Markets

No "price premium" for high quality workers

In e-commerce markets, sellers with a good reputation can sell their goods at a relatively high price (premium)

In crowdsourcing, the requester sets the price, and this is typically the same for all workers
Quality Control
part 3
Different Mechanisms for Quality Control

Aggregation and redundancy
Embedded gold standard data
Economic incentives
Reputation systems

Statistical models
Expectation Maximization algorithm

EM is an algorithm for finding the probabilities of unobserved variables. We will use it to estimate how accurate workers’ labels are, and infer how good each worker is. This is more sophisticated than voting.
Dawid and Skene (1977)

**Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm**

Examined application to medical diagnosis

Patients are sometimes treated by multiple physicians, who can give different diagnoses

Why? Doctors may have different questions. Patients may describe history differently. Doctors may classify symptoms differently
Observer Error

Given that different doctors have different opinions, they can't all be right. How often do individual physicians suffer from “observer error”? Are their errors systematic? Answers depend on the “true” diagnosis.
Observer Error

Observer error would be easy to calculate if we had ground truth
Simply count the misdiagnoses and divide by the total number of diagnoses
However, sometimes it is impossible to know what diagnosis is correct. Same set of symptoms can arise from multiple root causes.
“I know it when I see it”

I shall not today attempt further to define "hard-core pornography"; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it.

—Justice Potter Stewart
<table>
<thead>
<tr>
<th>url</th>
<th>worker 1</th>
<th>worker 2</th>
<th>worker 3</th>
<th>worker 4</th>
<th>worker 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun.com</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google.com</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>
Can't have Justice Stewart rule on everything
Instead, we will apply Dawid and Skene’s EM algorithm, which iteratively
1. Estimates the correct answers, using labels from multiple workers, and accounts for the quality of each worker
2. Estimates the quality of the workers by comparing the submitted answers to the inferred correct answers
Inputs

a set of $N$ objects $o_1 \ldots o_N$
sunnyfun.com, sex-mission.com, google.com, youporn.com, yahoo.com

a set of $L$ possible labels:
{porn, not porn}

Labels for each object by $K$ workers
worker1, worker2, worker3, worker4, worker5
Goal 1

Recover the true class label $T(o_n)$ for each object $o_n$ when “gold” truth is unknown. Since the true labels are not known / never directly observed, they are called latent variables.
Goal 2

For each worker who contributed labels, calculate their accuracy or reliability.
To calculate accuracy show how often they mistakenly choose one label when a different one is the actual truth.
Chicken and egg problem

If we knew what the true class labels were for each object for each object, then we could compute each Turker’s accuracy.

If we had accuracies for every Turker, then we could infer what the true label for each object should be.
**Input:** Labels $l[k][n]$ from worker $(k)$ to object $o_n$,

**Output:** Confusion matrix $\pi_{i,j}^{(k)}$ for each worker $(k)$, Correct labels $T(o_n)$ for each object $o_n$, Class priors $Pr\{C\}$ for each class $C$

1. Initialize error rates $\pi_{i,j}^{(k)}$ for each worker $(k)$ (e.g., assume each worker is perfect);
2. Initialize correct label for each object $T(o_n)$ (e.g., using majority vote);
3. while not converged do
   4. Estimate the correct label $T(o_n)$ for each object, using the labels $l[\cdot][n]$ assigned to $o_n$ by workers, weighting the votes using the error rates $\pi_{i,j}^{(k)}$;
   5. Estimate the error rates $\pi_{i,j}^{(k)}$, for each worker $(k)$, using the correct labels $T(o_n)$ and the assigned labels $l[k][n]$;
   6. Estimate the class priors $Pr\{C\}$, for each class $C$;
4. end
5. return Estimated error rates $\pi_{i,j}^{(k)}$, Estimated correct labels $T(o_n)$, Estimated class priors $Pr\{C\}$

**Algorithm 1:** The EM algorithm for worker quality estimation.
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun.com</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google.com</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>url</td>
<td>True Labels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sunnyfun.com</td>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex-mission.com</td>
<td>porn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>google.com</td>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>youporn.com</td>
<td>porn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yahoo.com</td>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Repeat until convergence

You can continue to iterate until your values converge
For this example, we converge after the first iteration
<table>
<thead>
<tr>
<th>Domain</th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun.com</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google.com</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn.com</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td></td>
<td>worker1</td>
<td>worker2</td>
<td>worker3</td>
<td>worker4</td>
<td>worker5</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>sex-mission</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>google</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>youporn</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>yahoo</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>worker2</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>porn</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>worker3</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>porn</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>worker4</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>porn</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>worker5</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>porn</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>worker1</td>
<td>worker2</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>sex-mission</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>google</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>youporn</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>yahoo</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Initialize confusion matrices to uniform
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

Compute labels using majority vote

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>sex-mission</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>google</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>youporn</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>yahoo</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>worker1</td>
<td>worker2</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>sex-mission</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>google</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>youporn</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>yahoo</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Compute labels using majority vote
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sex-mission</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>google</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>youporn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>yahoo</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Treat the correct label as the one with the most votes.
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

Recompute worker confusion matrices...as if though your labels are 100% correct
Recompute worker confusion matrices... as though your labels are 100% correct.
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

Recompute worker confusion matrices...as though your labels are 100% correct
Recompute worker confusion matrices...as though your labels are 100% correct.
Renormalize confusion matrices (based on true labels)

<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sex-mission</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>google</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>youporn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>yahoo</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>worker1</td>
<td>worker2</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sex-mission</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>google</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>youporn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>yahoo</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>worker1</td>
<td>worker2</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
</tr>
</tbody>
</table>

Recompute labels using weighted majority vote.

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>sex-mission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>google</td>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td>youporn</td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td>yahoo</td>
<td></td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>worker1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>worker2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>worker3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>worker4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>worker5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>not</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>worker1</td>
<td>worker2</td>
<td>worker3</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
</tr>
</tbody>
</table>

- **Recompute labels using weighted majority vote**

<table>
<thead>
<tr>
<th>worker1</th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>porn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0.67</td>
<td>0.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>worker2</th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>porn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0.33</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>worker3</th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>porn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>worker4</th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>porn</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>not</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>worker5</th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>porn</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>not</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>worker1</td>
<td>worker2</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
</tr>
</tbody>
</table>

Recompute labels using weighted majority vote.
Treat the correct label as the one with the most votes.
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

Recompute worker confusion matrices...as though your labels are 100% correct.
<table>
<thead>
<tr>
<th></th>
<th>worker1</th>
<th>worker2</th>
<th>worker3</th>
<th>worker4</th>
<th>worker5</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>sex-mission</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
</tr>
<tr>
<td>google</td>
<td>not</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
<tr>
<td>youporn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>porn</td>
<td>not</td>
</tr>
<tr>
<td>yahoo</td>
<td>porn</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>porn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>porn</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunnyfun</td>
<td>0.26</td>
<td>0.74</td>
</tr>
<tr>
<td>sex-mission</td>
<td>0.69</td>
<td>0.31</td>
</tr>
<tr>
<td>google</td>
<td>0.29</td>
<td>0.71</td>
</tr>
<tr>
<td>youporn</td>
<td>0.82</td>
<td>0.18</td>
</tr>
<tr>
<td>yahoo</td>
<td>0.26</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Iterate until convergence.
Question

How would you use gold standard data in the EM process?
EM Algorithm

Re-Calculate Worker Scores over two steps:

1. Estimate the probability that each answer is correct, using labels from multiple workers weighted by the probability that they are correct
2. Estimate the quality of the workers by comparing their submitted answers to the inferred correct answers
Confusion Matrix Gives us Worker Error

From the confusion matrix we can measure the overall error rate for each worker. Sum of the non-diagonal elements of the confusion matrix (weighted by the priors) This results in a single, scalar value as the quality score for each worker.
Worker error

<table>
<thead>
<tr>
<th>Worker</th>
<th>Porn</th>
<th>Not</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>Worker2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>Worker3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Worker4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Worker5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Worker1: 0.67
Worker2: 0.33
Worker5: 1.5
Advanced Topics

Bias versus error
How noisy can the workers be and still allow us to still converge to a correct solution?
Bias versus error

Error rate alone is not sufficient to measure the inherent value of a worker. For example, workers may be careful but biased. In a non-binary case, this is more apparent.

What if instead of asking our workers to label sites porn or not porn, we asked them to label the G, PG, R, X?
Bias versus error

Parents with young children tend to be more conservative
They tend to classify PG-rated sites as R-rated sites, and R-rated sites as X-rated.
Such workers give consistently and predictably incorrect answers
It is possible to automatically correct for bias
Implications

Unlike with spammers, with biased workers it is possible to “reverse” the errors
We can recover a label assignment of much higher quality
In the presence of systematic bias, the naive measurement of error rate results in underestimates of the true quality of the worker
This potentially leads to incorrect rejections and blocks of legitimate workers
For more details

Check out two papers by Panos Ipeirotis and his collaborators

Managing Crowdsourcing Workers discusses separating error and bias

Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers discusses how noisy judgements can be, with us still getting good quality results