Social Computing Systems

Walter S. Lasecki

EECS 498/598, Winter 2018
(http://tiny.cc/socsClass)
Today

- Crowd-powered systems
- The need for quality control in crowd-powered systems
- Quality control methods in crowdsourcing
- Workflows
So far...

- Human computation can solve problems that automatic methods cannot

- Crowdsourcing can support flexible, highly-available human computation
  - Platforms like MTurk and Upwork provide APIs for hiring

- This can improve the state of work

- We want to be able to use all of this to create new, more powerful systems
Quality Control
The world’s largest community of biased coins collectors...

1,000 participants on Amazon Mechanical Turk flip a coin and report “h” (heads) or “t” (tails)

- 65% heads
- 28% tails
- 7% head, heads, Tail, tails, Tails, t for tails, talis, f

Number of Responses
How many M&Ms do you think are in this Pittsburgh Penguins container?

A green one is shown next to the container to help with scale. They’re Peanut M&Ms.

It’s for science. Thanks!
Take a guess here:

http://shoutkey.com/quince
Quality Considerations

- Confidence
- Accuracy
  - Precision
  - Recall
- Discoveries
  - Number
  - Quality
- [...]

[31x336]Quality Considerations

- Confidence
- Accuracy
  - Precision
  - Recall
- Discoveries
  - Number
  - Quality
- [...]

[72x91]…
Basic Approaches

- **Averaging**
  - Works well when answers are independent and identically distributed (i.i.d)
  - Works well with larger populations
  - Does not work well with questions likely to result in biased answers

- **Median**
  - Good in cases where only *most* answers are i.i.d
  - More stable to radical answers than averaging

- **Mode / majority voting**
  - More stable to radical answers than averaging
  - Works even if bias exists, but only if the bulk of workers know the answer
Classes of Questions

- Based on the crowd’s knowledge
  - Class 1: Everyone in the crowd knows the answer
    - Maybe it requires varying amount of effort, but almost everyone *could* know the answer
Classes of Questions

- Based on the crowd’s knowledge
  - Class 1: Everyone in the crowd knows the answer
    - Maybe it requires varying amount of effort, but almost everyone *could* know the answer
  - Class 2: At least some in the crowd know the answer
    - This could be one expert, or more than one
    - May not be a majority of workers, but majority can recognize the answer
Classes of Questions

- Based on the crowd’s knowledge

- Class 1: Everyone in the crowd knows the answer
  - Maybe it requires varying amount of effort, but almost everyone *could* know the answer

- Class 2: At least some in the crowd know the answer
  - This could be one expert, or more than one
  - May not be a majority of workers, but majority can recognize the answer

- Class 3*: Only the collective knows the answer
  - No individual knows the answer
Classes of Questions

- Based on the crowd’s knowledge
  - Class 1: Everyone in the crowd knows the answer
    - Maybe it requires varying amount of effort, but almost everyone *could* know the answer
  - Class 2: At least some in the crowd know the answer
    - This could be one expert, or more than one
    - May not be a majority of workers, **but majority can recognize the answer**
  - Class 3*: Only the collective knows the answer
    - No individual knows the answer
  - Class 4*: No one in the crowd knows the answer
    - Tricky class of problems… so solve them, we need to convert them to class-2 or 3 problems
Motivating Workers

- Given the challenges of aggregation, we want to make sure everyone tries
- How do we do this?
- How do the efforts of others affect each worker?
- Can we generalize how we think about these factors?
How do we overcome Class 4 problems?

- Modify some of our assumptions

- What are our assumptions?
  - Workers all have uniform expertise (as far as we can tell)
  - All answers are equally likely
  - Workers all put in the same effort
  - Answers all have a single step
  - ...
Gold-Standard Questions

- Questions that we know the answers to
- Selected to be predictive of a skill, or to show general accuracy
- Could demonstrate willingness to put in certain levels of effort

Creating gold questions can be expensive
- Manual creation of an example set
- What % gold is needed? …too low and workers might assume they don’t exist
- Does not scale well if variety is needed, or there are many classes of problem

Crowdsourcing gold creation
- Consensus/collective-intelligence mechanisms
- Potentially faulting (as discussed before), but can lead to more affordable gold
Answer Likelihoods

- How likely are two workers to agree?
  - Multiple choice?
  - Open-ended text?
  - Image?
  - UI controls?

- Is there any bias?
  - Recall the Legion robot example:
    - The wrong feedback leads to over-agreement
    - Over-agreement can result in consensus mistakes
  - Saliency leads to bias
    - M&Ms: 0, 1000, 1000000, …
    - Scribe (captioning): first word in sentence
Predicting Effort

- Levels of effort
  - Lazy worker (-)
  - ‘Average’ worker (+)
  - Eager beaver (-)

- Gold standard questions can help assess
  - Does not catch all classes by default

- Other approaches?
Refining Answers

- Let workers build on one another’s work
  - Collaborate? (provide justification, let people query each other, etc.)
  - Anonymous revision (one answer can be revised / replaced with the next)

- Risks
  - ‘Groupthink’ / shared mistakes
  - Destructive behavior
    - Add noise
    - Shredder challenge…
Shredder Challenge (DARPA)

- After the crowd’s success in the Red Balloon grand challenge, DARPA launched the Shredder grand challenge
  - Reconstruct a shredded document
  - Number of pieces ranged from hundreds to thousands
  - Computers are not able to do this for large sets. Can the crowd?

- Several teams entered the competition
  - UCSD team entered using an open platform that let anyone help out
How a Lone Hacker Shredded the Myth of Crowdsourcing

High-tech analysis of a 2011 DARPA Challenge shows why we can’t have nice things

(Medium article)
This is the previously untold story of how and why Adam humbled some of the brightest brains in computer science, their years-long search to find him, and the researchers who now believe that the wisdom of the crowd might be nothing more than a seductive illusion.

How a Lone Hacker Shredded the Myth of Crowdsourcing

High-tech analysis of a 2011 DARPA Challenge shows why we can’t have nice things

(Medium article)
What went wrong?

- Open platform, but not robust
  - Allowed for disproportionate destruction per worker effort

- Didn’t account for malicious workers (at all)

- In the end: 1-2 adversarial workers killed the effort
  - (...when even a wave of 4chan attackers got bored)
  - At least 1 used a Sybil attack
  - Stacked chads
  - Moved chads off working area/canvas
  - Used multi-select (drag-to-select)
  - Demoralized genuine workers

[artist's depiction]
(well, my depiction, anyway)
Workflows

EECS 498/598
SoCS — Winter 2018
Why workflows?

The wave of the future!
Why workflows?

The wave of the future!

...and the present

“Illustration showing Human Task. It shows a BPEL process in JDeveloper connected to Workflow Services Human Workflow.”

Oracle.com
Why workflows?

The wave of the future!

...and the present

...and the past
Why workflows?

Another way to overcome “Class 4” crowdsourcing problems

Why? How?

- Task decomposition limits complexity, both of the individual task and overall
- This lets us use simple tasks with complex interdependence that workers do not have to deal with themselves (so the overall task can be complex)
What is a Workflow?

“The sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion.” — Google

... a human (+machine) “algorithm”.

Much like in code, these algorithms let us recombine simple steps to solve complex problems.
Workflow (Google Image Search Result)
Unit of human computation

t := 0;
initialize: \( P(0) := \{ \tilde{a}_1(0), ..., \tilde{a}_\mu(0) \} \in I^\mu \);
evaluate: \( P(0) : \{ \mathcal{A}(\tilde{a}_1(0)), ..., \mathcal{A}(\tilde{a}_\mu(0)) \} \);

while not \( P(t) \) do
    select: \( P'(t) := s_{\mathfrak{a}_1}(P(t)) \);
    recombin: \( P''(t) := \otimes_{\mathfrak{a}_r}(P'(t)) \);
    mutate: \( P''(t) := m_{\mathfrak{a}_m}(P''(t)) \);
    evaluate: \( P'''(t) : \{ \mathcal{A}(\tilde{a}_1'''(t)), ..., \mathcal{A}(\tilde{a}_\mu'''(t)) \} \)
    replace: \( P(t + 1) := r_{\mathfrak{a}_r}(P'''(t) \cup Q) \);
<communication step>
\( t := t + 1 \);
end while
Traditional Workflow
(what we’ve seen so far)
Iterative Improvement & Voting

Turkit & Sanjay Kairam
Handwriting Recognition

- **Final Version:**
  - You (misspelled) (several) (words). Please spellcheck your work next time. I also notice a few grammatical mistakes. Overall your writing style is a bit too *phoney*. You do make some good (points), but they got lost amidst the (writing). (signature)
Soylent is a prototype crowdsourced word processing interface. It focuses on three main tasks: shortening the user's writing, proofreading [...]
CrowdForge

Figure 1: Overview of framework for splitting up and recombining complex human computation tasks.

Ask most people who plan to travel to New York City what they want to see while they are there and invariably you will hear about the top tourist attractions: the Empire State Building, the Statue of Liberty, and the Grand Central Terminal, with the Empire State Building probably coming in as number one on the list of "must see" for visitors to the city. No wonder: the Empire State Building has a long history, having celebrated its seventy-fifth anniversary on May 1, 2006. Yet the Statue of Liberty is also a popular tourist destination.
Figure 2: The PlateMate system. Work travels between stages and Human Intelligence Tasks (HITs) along the black arrows, starting from the input on the left and concluding with the output on the right. The system takes submitted photos and creates Tag tasks to annotate these photos with boxes. Each box becomes the input to a series of Identify tasks which end with a list of foods from a commercial food database. Each individual food is then input to a Measure task, which produces a unit and amount. Dashed boxes represent optional stages, which may be skipped during routing.
Crowdsourcing Crowd Workflows?

Turkomatic: Automatic, Recursive Task and Workflow Design for Mechanical Turk
So, we can do anything now…
right?
So, we can do anything now…

right?

… right??
So, we can do anything now…
right?

… right??

Sort of!
Back to the high level view...
Task Decomposition
Task Decomposition
Task Decomposition
Task Decomposition
How much help is too much?
Example: Task Decomposition

Hye Fred,

It was great talking with you the other day! We should definitely try to get that proposal submitted soon,

Thanks Sally!

- J.D. Officeworker
Hye Fred,

It was great talking with you the other day! We should definitely try to get that proposal submitted soon.

Thanks Sally!

- J.D. Officeworker
Example (v1): Worker 1

Hye Fred,

It was great talking with you the other day!
Example (v1): Worker 2

We should definitely try to get that proposal submitted soon.

Thanks Sally!

- J.D. Officeworker
We should definitely try to get that proposal submitted soon.

Thanks Sally!

- J.D. Officeworker
Example (v2): Worker 1

Hye Fred,
Example (v2): Worker 2

It was great.
Example (v2): Worker 3

talking with you
Example (v2): Worker 4
Example (v2): Worker 5

Otter day!
Example (v2): Worker 6

We should
Example (v2): Worker 7
definitely trie
Example (v2): Worker 8
to get that proposal
Example (v2): Worker 8

submitted soon,
Example (v2): Worker 10

Thanks Sally!

- J.D. Officeworker
Example (v2): Workers

Hye Fred,

It was great talking with you the other day! We should definitely try to get that proposal submitted soon,

Thanks Sally!

- J.D. Officeworker