Crowdsourcing: Opportunities and Challenges

Deepak Ganesan
Associate Professor
UMass Amherst

Rise of Crowdsourcing

Crowdsourcing = Harvesting society’s wisdom, skill, creativity, and scale to solve a task

Examples of crowdsourcing systems

- Research
- Logo
- Funding
- Engineering
- SEO/SEM
- Lead-Gen
- Sales
- Customer Support
- R&D
- CSR
- Photos
- Personalized Songs
- Voice-overs
- Image Manipulation
- Translation
- Photo Tagging
- Transcription
- Surveillance Analytics
Examples of crowdsourcing systems

<table>
<thead>
<tr>
<th>Company</th>
<th>Initiative</th>
<th>Goal</th>
<th>Community Size</th>
<th>Countries Spanned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Mechanical Turk</td>
<td>general-purpose</td>
<td>300,000 e</td>
<td>34 *</td>
</tr>
<tr>
<td>Burda Media</td>
<td>BurdaStyle</td>
<td>fashion design</td>
<td>298,016</td>
<td>46 **</td>
</tr>
<tr>
<td>CNN</td>
<td>iReport</td>
<td>news reporting</td>
<td>452,536</td>
<td>undisclosed</td>
</tr>
<tr>
<td>Eli Lilly</td>
<td>InnoCentive</td>
<td>scientific R&amp;D</td>
<td>160,000</td>
<td>175</td>
</tr>
<tr>
<td>Facebook</td>
<td>Translations</td>
<td>language translation</td>
<td>300,000</td>
<td>180</td>
</tr>
<tr>
<td>Google</td>
<td>Image Labeler</td>
<td>image labeling</td>
<td>undisclosed</td>
<td>undisclosed</td>
</tr>
<tr>
<td>Netflix</td>
<td>Netflix Prize</td>
<td>data mining</td>
<td>41,305</td>
<td>186</td>
</tr>
</tbody>
</table>

Classifying by complexity

Outline

- What is crowdsourcing?
- Human computation using Mechanical Turk
- Crowdsourced data collection using mCrowd
- Course Overview
- Project Ideas

What is the Amazon Mechanical Turk?
AMT Basics

- Mechanical Turk provides a set of primitives
  - HIT properties (reward, instructions, requirements, qualifications)
  - Assignment can be approved or rejected
  - Worker can be bonused
  - Worker can be blocked
  - Qualification can be assigned or revoked

- No explicit requester reputation but several websites (e.g. turkopticon) provide information on good/bad requesters.

AMT as a Research Enabler

Why is AMT popular for research?

- **Scalable**: 50K+ humans in steady state.
- **Fast**: Rapid responses from thousands of workers.
- **Cheap**: One or few cents per task.
- **Hassle-free**: Subject anonymity/identifiability/pre-screening/diversity

Challenges in using AMT

- How much to pay workers?
- How to reduce delay for responses?
- How to maximize accuracy?
How much to pay workers?

How to reduce delay?

Outline

- What is crowdsourcing?
- Human computation using Mechanical Turk
- Crowdsourced data collection using mCrowd
- Course Overview
- Project Ideas

Mobile crowdsourcing for data creation

leverage millions of smartphone users to provide real-time, context-aware data about environment, transportation, health, civic issues, etc
Rise in Mobile Crowdsourcing Apps

mCrowd: A Task Market for Mobile Sensing

mCrowd Architecture

mCrowd: Viewing Tasks
**mCrowd: Creating a Task**

- Specify widgets
- Geo-scope
- Specifying widgets
- Geo-scope
- Incentives
- Blacklist/Whitelist
- Deadline
- Data publishing

**Why mCrowd?**

- Enable micro-crowdsourcing efforts
- Avoid fragmented user base
- Explore diverse incentives for user retention

**Why mCrowd?**

- Enable micro-crowdsourcing efforts
- Gather water quality info on a stream near home.
- Setup crowdsourcing system for students on a field trip.

---

**Why mCrowd?**

- Enable micro-crowdsourcing efforts
- Avoid fragmented user base
- Explore diverse incentives for user retention

**Percentage Retention**

- 100%
- 75%
- 50%
- 25%

**Reward:**

- Points
- $$ rewards
- Time-varying incentives
- Location-based incentives

---

**Why mCrowd?**

- Enable micro-crowdsourcing efforts
- Avoid fragmented user base
- Explore diverse incentives for user retention

**Retention Over Time**

- 0 days
- 30 days
- 60 days
- 90 days

**Flurry Analytics**

---

**Why mCrowd?**

- Enable micro-crowdsourcing efforts
- Avoid fragmented user base
- Explore diverse incentives for user retention

**Retention Over Time**

- 0 days
- 30 days
- 60 days
- 90 days

**Flurry Analytics**
Why mCrowd?

- Enable micro-crowdsourcing efforts
- Avoid fragmented user base
- Explore diverse incentives for user retention
- Study data quality assessment issues

Outline

- What is crowdsourcing?
- Human computation using Mechanical Turk
- Crowdsourced data collection using mCrowd
- Course Overview
- Project Ideas

What do I expect from you?

- Taking course for one credit:
  - Two paper presentations
  - Written reviews for any ten papers
- Taking course for three credits:
  - Two paper presentations
  - Written reviews for any twenty papers
  - Significant course project
- All reviews will be online on the course webpage.

Course structure

- Invited Talks
  - Nathan Eagle (MIT) - Mobile crowdsourcing
  - Panos Ipeirotis (NYU) - Data quality management
  - Arvind Thiagarajan (MIT) - Traffic crowdsourcing
  - Jordan Boyd-Graber (UMD) - NLP crowdsourcing
  - Chris Callison-Burch (JHU) - NLP crowdsourcing
  - John Horton (Harvard) - Policy/Economics & crow...
  - Rob Baker (Ushahidi) - Disaster relief & management
  - Shaili Jain (Yale) - Incentives in crowdsourcing
  - Murat Demirbas (UBuffalo) - Twitter for sensing..
Course structure

- Papers from several conferences/workshops
  - NAACL, MobiSys, Mobicom, Ubicomp, CHI, EC, AAAI, KDD, SenSys, HCOMP ...

- Need four volunteers for papers next week:
  - Games with a purpose (IEEE Computer)
  - Predicting the present with Google Trends (Google TR)
  - TurKit: Human Computation Algorithms on ... (UIST 2010)
  - Who are the crowdworkers? Shifting demogra...(CHI 2010)

Paper presentations

- Paper presentation: 15 minutes
  - Discuss the main ideas in the paper.
  - Focus on new aspects that have not been discussed earlier in the seminar.

- End with discussion points: you are responsible for leading a discussion on the paper.

- 10 minute discussion of the paper.

Outline

- What is crowdsourcing?
- Human computation using Mechanical Turk
- Crowdsourced data collection using mCrowd
- Course Overview
- Project Ideas

Available Software for AMT Projects

- get-another-label: A set of tools for managing redundant answers collected from Amazon Mechanical Turk
- turkit: Tools for iterative tasks on Mechanical Turk
- CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones
  - http://data.doloreslabs.com
- Mechanical Turk Monitor: General Data
Project Idea 1: Using AMT to process data

- Pick a data corpus that is “hard” to process using automated computer algorithms.
  - Should improve on existing approaches, and/or demonstrate a new application domain for AMT.
  - Example: Data cleaning engine for sensor data
    - Accelerometer, GPS traces, temperature traces, ..etc
    - Can AMT workers help us make better sense of this data than ML algorithms to detect events?

Project Idea 2: Toolkit to Optimize Delay, Quality, Cost

- Existing toolkits focus on data quality. Design toolkit that offers delay-quality-cost tradeoffs in using AMT.
- Model the online behavior of Turkers for the task:
  - time-of-day effects
  - price vs delay behavior
  - data quality for individuals.
- Use learnt behavior to iteratively improve performance over time.

Project Idea 3: Crowdsourced Measurement Infrastructure

- Use AMT + mCrowd to augment existing PlanetLab-based internet measurement infrastructure.
  - Internet measurement is largely based on using fixed infrastructure (e.g. iPlane). But large swaths of the Internet are not measured using this framework.
  - Can we utilize crowdsourcing to augment existing infrastructure?

Project Idea 4: Fine-grained model of AMT

- Monitor behavior of AMT at fine-granularity (minutes) using PlanetLab. Validate power-law distribution and other conclusions from existing studies.
**Project Idea 5: MapReduce for AMT Tasks**

- Design a MapReduce-like programming framework for using AMT
- **Commonalities**
  - Large task divided into smaller sub-problems
  - Work distributed among worker nodes (turkers)
  - Collect all answers and combine them
- **Challenges:**
  - Delay variability in obtaining responses
  - Use Turker reputation to determine map function
  - Support for delay-cost-quality tradeoffs

*(based on idea from Omar Alonso, Microsoft)*

**Available software for mobile crowdsourcing**

- **Ushahidi**
  - Crisis crowdsourcing: Support for several phones, web-backend, report validation engine, Mechanical Turk filtering engine...
  - mCrowd: support for iPhone; surveys, image, audio collection; $$ + points incentives; REST APIs; Web backend (contact: Moaj Musthag)
- **RCP**
  - Crowdsourced data collection campaigns. Web backend; Android code for instances.
- **opendatakit**
  - Open Data Kit is a suite of tools to help organizations collect, aggregate and visualize their data.

**Project Idea 6: Data Quality filtering engine using mCrowd + AMT**

- Use Crowdsourcing for data quality assessment
- **Redundancy**  **Bad data**  **Wrong species**

**Image processing** → **Filter using untrained masses** → **Expert validation**

**Project Idea 7: A Privacy Engine for mCrowd**

- Privacy is an important problem when dealing with data from phones.
- Design a privacy engine for mCrowd that:
  - provides simple & effective policies for data provider
  - supports backend data anonymization/perturbation/..
Project Idea 8: Have an application in mind?

• Design a mobile crowdsourcing application for your favorite cause.

• Given the time constraints of course:
  • keep development time small (1.5 months)
  • focus on deployment study with a reasonable number of users.

Project Idea 9: Come up with something that excites you...