Data CrowdSourcing: Is It For Real?

Hector Garcia-Molina

(work with Steven Whang, Peter Lofgren, Aditya Parameswaran, Hyunjung Park, Vasilis Verroios, Manas Joglekar, Ming Han Teh and others)

Stanford University
Man/Woman vs. Machine
John Connor

>  

SkyNet Terminator
CrowdSourcing

“performing a task using human workers that solve sub-problems”
CrowdSourcing

REWARD
($5,000.00)

Reward for the capture, dead or alive, of Wm. Wright, better known as
“BILLY THE KID”

Age, 18. Height, 6 feet, 3 inches. Weight, 125 lbs. Light hair, blue
eyes and even features. He is the leader of the worst band of
desperadoes the Territory has ever had to deal with. The above
reward will be paid for his capture or positive proof of his death.

JIM DALTON, Sheriff.

DEAD OR ALIVE!
“BILLY THE KID”
CrowdSourcing
Crowd Sourcing

Real-time, human, copy-editing and proofreading for everything you write

Wordy is a professional copy-editing and proofreading service. We optimise the accuracy and readability of content – from Fortune 500 business reports to academic texts and web copy.

Submit your order and let Wordy find the best editor for the job. Queries are resolved in real time, while turnaround times average 1,200 words per hour.

Process all major file formats and Google Docs

We track the changes of everything we do
Crowd Funding
Crowd Funding
More CrowdSourcing Examples

Categorizing Images

Search Relevance

Data Gathering

Image Matching Translation
CrowdSourcing: Final Examples
CrowdSourcing: Spammers & Porno

- Coordinator
  - Open account: Email service
  - Solve: CAPTCHA
  - Answer:Porno site
CrowdSourcing
CrowdSourcing Space
CrowdSourcing Space

coordinator

human

workers

computer

computer

human
CrowdSourcing Space

coordinator
distributed computing
workers

1 old-fashioned

2

3 current focus
CrowdSourcing Space

cordinator

human

workers

computer

human

data

money

document

training set

output
CrowdSourcing Space

computer
coordinator
human

workers
data

money
document
training set

output

this talk
Two More Dimensions

incentives

money

hidden

philanthropy

entertainment

small (micro)

large

granularity
So, Is CrowdSourcing for Real??

- Is it used in practice?
- Are there interesting research problems?
Many Crowdsourcing Marketplaces!
Many Research Projects!
Overview: Crowd Data Management

- Data Processing
- Data Gathering
- Searching
Finding the Maximum

What is the best applicant for the job?
Fundamental Tradeoffs

Latency

Cost

Uncertainty
Fundamental Tradeoffs

Latency

Cost

Uncertainty

focus on these for now
Example Max Algorithm FS
Example Max Algorithm FP
Latency-Cost Tradeoff

- time
- cost

FS

FP

?
Latency-Cost Tradeoff

![Diagram of Latency-Cost Tradeoff]

- **FS**
- **Budget**
- **FP**

- time
- cost

?
Latency-Cost Tradeoff

Min-Latency Problem

Budget

Smallest time

FS

FP

cost
time
Framework

• Input:
  – Question budget \( b \)
  – Number of elements \( c \)
  – Latency function \( L(q) \): time to answer \( q \) questions

• Reliable workers (use Reliable Worker Layer)

• Proceed in rounds

• First, select budget/round, e.g., \( (10, 7, 7, 5) \)

• Then use Question Selection Algorithm in each round
Framework

- **Input:**
  - Question budget $b$
  - Number of elements $c$
  - Latency function $L(q)$: time to answer $q$ questions

- **Reliable workers (use Reliable Worker Layer)**
- **Proceed in rounds**
- **First, select budget/round, e.g., (10, 7, 7, 5)**
- **Then use Question Selection Algorithm in each round**
Examples

budget \( b = 30 \); elements \( c = 12 \)

budget vector\( = (15, 10, 1) \)

budget vector\( = (6, 6, 1) \)
Examples

budget = 30; elements = 12

budget vector = (15, 10, 1)

budget vector = (6, 6, 1)

Round 1

Round 2

Round 3

Question Selection:
Tournament Graphs
Examples

budget = 30; elements = 12

budget vector = (15, 10, 1)

budget vector = (6, 6, 1)

num. groups \Rightarrow remaining elements
"memoryless"

Question Selection:
Tournament Graphs
Focus on Tournament Graphs

• How to find optimal budget vector?
• Example, which is best for $b=70$, $c=40$
Focus on Tournament Graphs

- How to find optimal budget vector?
- Example, which is best for $b=70$, $c=40$

Note: Goal is not minimum questions, but minimum latency
tDP Algorithm

- Assuming tournament graph question selection, our tDP Algorithm finds optimal budget vector
- Can use dynamic programming because of nice properties of tournament graphs
How Does tDP Work (optional slide)

<table>
<thead>
<tr>
<th>budget, b</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

number of elements, c

**Optimal latency**

fill in table
How Does tDP Work (optional slide)

<table>
<thead>
<tr>
<th>budget, b</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- say we start with first round that reduces elements 6->3;
- this tournament costs 3 questions, remaining 5-3=2
- \( y = L(3) + x \)
How Does tDP Work (optional slide)

- consider all possible reductions for first round
- pick for $y$ one that yields minimum latency

<table>
<thead>
<tr>
<th>budget, $b$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

number of elements, $c$
BUT wait, there is more!!

- tDP + tournament graphs has better (lower) worst case latency than any budget allocation scheme with any question selection algorithm!
BUT wait, there is more!!

- tDP + tournament graphs has better (lower) worst case latency than any budget allocation scheme with any question selection algorithm!
- And in practice, tDP + tournament graphs is "damn good" for average case latency (see experiments)
Key Insight

worst case remaining elements: \{a,c,e,g,i,k\}

same outcome but better latency!
Example of Experimental Results

budget $b=4000$, all using tournament graphs
Beyond Max

- Filtering
- Sorting
- Clustering
- Entity Resolution
- Adding terms to a taxonomy
- Building a Folksonomy
- ...

48
Overview: Crowd Data Management

• Data Processing
• Data Gathering
• Searching
Crowd As Information Source

Declarative queries

DBMS like thing

Web
Example #3: DeCo

*Work with Aditya Parameswaran, Hyunjung Park, Jennifer Widom
## Small Example

<table>
<thead>
<tr>
<th>restaurant</th>
<th>rating</th>
<th>cuisine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>French</td>
</tr>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>California</td>
</tr>
<tr>
<td>Bytes</td>
<td>3.8</td>
<td>California</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

User view
Small Example

<table>
<thead>
<tr>
<th>restaurant</th>
<th>rating</th>
<th>cuisine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>French</td>
</tr>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>California</td>
</tr>
<tr>
<td>Bytes</td>
<td>3.8</td>
<td>California</td>
</tr>
</tbody>
</table>

User view

Anchor

Dependent

Dependent
# Small Example

<table>
<thead>
<tr>
<th>restaurant</th>
<th>rating</th>
<th>cuisine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>French</td>
</tr>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>California</td>
</tr>
<tr>
<td>Bytes</td>
<td>3.8</td>
<td>California</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **User view**

- **Anchor**

- **Dependent**

- **fetch rule**
Small Example

<table>
<thead>
<tr>
<th>restaurant</th>
<th>rating</th>
<th>cuisine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>French</td>
</tr>
<tr>
<td>Chez Panisse</td>
<td>4.9</td>
<td>California</td>
</tr>
<tr>
<td>Bytes</td>
<td>3.8</td>
<td>California</td>
</tr>
</tbody>
</table>

User view

Anchor

Dependent

Resolution rule

Dependent
Example #4: CorwdFill*

- Goal: Collect high-quality structured data from the crowd, while capping total monetary cost and keeping latency low

<table>
<thead>
<tr>
<th>name</th>
<th>nationality</th>
<th>position</th>
<th>caps</th>
<th>goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Messi</td>
<td></td>
<td>FW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klose</td>
<td>Germany</td>
<td></td>
<td>133</td>
<td></td>
</tr>
</tbody>
</table>

*Hyunjung Park
**CrowdFill Prototype**

<table>
<thead>
<tr>
<th>name</th>
<th>nationality</th>
<th>position</th>
<th>caps</th>
<th>goals</th>
<th>thumbs_up</th>
<th>thumbs_down</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lionel Messi</td>
<td>Argentina</td>
<td>FW</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ronaldinho</td>
<td>Brazil</td>
<td>MF</td>
<td></td>
<td>Empty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neymar</td>
<td>Brazil</td>
<td>FW</td>
<td></td>
<td>Empty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iker Casillas</td>
<td>Spain</td>
<td>FW</td>
<td>150</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ronaldinho</td>
<td>Brazil</td>
<td>FW</td>
<td></td>
<td>Empty</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview: Crowd Data Management

- Data Processing
- Data Gathering
- Searching
Example #5: DataSift

• Can Your Search Engine Handle This?

buildings in the vicinity of

type of cable that connects to

apartments in a good school district near Somerville, with a bus stop near by
<table>
<thead>
<tr>
<th>Rank</th>
<th>Product Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Mediabridge Hi-Speed USB 2.0 Cable - (6 Feet)</strong>&lt;br&gt;Product page: <a href="http://www.amazon.co/dp/B001MXLD4G">http://www.amazon.co/dp/B001MXLD4G</a>&lt;br&gt;Price: USD 4.99</td>
</tr>
<tr>
<td>5</td>
<td><strong>Mediabridge Hi-Speed USB 2.0 Cable - (10 Feet)</strong>&lt;br&gt;Product page: <a href="http://www.amazon.co/dp/B001MSU1HG">http://www.amazon.co/dp/B001MSU1HG</a>&lt;br&gt;Reviews: <a href="http://www.amazon.co/reviews/iframe?akid=AKIAJ...">http://www.amazon.co/reviews/iframe?akid=AKIAJ...</a>&lt;br&gt;Price: USD 5.49</td>
</tr>
</tbody>
</table>
DataSift* can handle rich queries!

*work with Aditya Parameswaran and Ming-Han Teh
DataSift Steps (One Way)

- Start with rich query Q
- Ask crowd for keyword queries \{K\} (and target)
- Run queries \{K\} and get some results \{D\}
- Ask crowd to evaluate \{D\} (w.r.t. Q)
- “Calibrate” queries \{K\}
- Get more results
Crowd Component 1: G (Gather)

Provide 3 possible distinct queries you would issue to Amazon Products to find products matching the description:

You may need to use your general knowledge to summarize the requirements into a suitable search phase. (See examples)
You should click the "Try Search" button to test your query.

Query 1: 
Query 2: 
Query 3: 

Example A: If the question asks for "point reyes; showing lighthouse only", your search query might be "point reyes lighthouse"
Example B: If the question asks for "SF bridge; night scene", your search query might be "golden gate bridge night"
Example C: If the question asks for "smartphone by apple", your search query might be "iphone" or "apple smartphone"
Crowd Component 2: F (Filter)

Which of the following products depicts the type of cable that connects to a printer?

<table>
<thead>
<tr>
<th>Thumbnail</th>
<th>Product Details</th>
<th>Rate</th>
</tr>
</thead>
</table>
| ![Cable](image) | **AmazonBasics USB 2.0 A-Male to B-Male Cable with Lighted Ends - Braided (6 Feet/1.8 Meters)**  
Product page: [http://www.amazon.co/dp/B003ES5ZQE](http://www.amazon.co/dp/B003ES5ZQE)  
Price: USD 6.99  
Description:  
6FT Braided USB A to B Cable with Light  
Features:  
- USB A(with blue LED)--B(with blue LED) Braided Cable  
- Ships in Certified Frustration-Free Packaging | ![Yes](image) ![No](image) |
| ![Cable](image) | **Mediabridge Hi-Speed USB 2.0 Cable - (10 Feet)**  
Product page: [http://www.amazon.co/dp/B001MSU1HG](http://www.amazon.co/dp/B001MSU1HG)  
Price: USD 5.49  
Description:  
**Mediabridge High-Speed USB 2.0 A-Male to B-Male Cable**  
* Full 480-Mbps Transmission Speed of The USB 2.0 Standard  
* Foil and Braid Shielding  
* Gold-Plated Copper Contacts  
* Limited 1 Year Warranty | ![Yes](image) ![No](image) |
The Elephant(s) in the Room...

- Crowds are Slow! Crowds are Costly!
- Want to use DataSift selectively!
Conclusion

• Is crowdsourcing for real??
  – YES!!

• Many interesting problems:
  – Crowd data processing
  – Crowd gathering
  – Search
  – Many others!
Top Secret

- Check in word of the day: "Palo Alto"
References

• Challenges in Data Crowdsourcing, IEEE Transactions on Knowledge and Data Engineering, 2016 (with Manas Joglekar, Adam Marcus, Aditya Parameswaran, Vasilis Verroios).


• An Overview of the Deco System: Data Model and Query Language; Query Processing and Optimization, SIGMOD Record, Volume 41, Dec 2012 (with Hyunjung Park, Richard Pang, Aditya Parameswaran, Neoklis Polyzotis, and Jennifer Widom).

• Datasift: An Expressive and Accurate Crowd-Powered Search Toolkit, 1st Conf. on Human Computation and Crowdsourcing (HCOMP), Palm Springs, USA, Nov 2013 (with Aditya Parameswaran, Ming Han Teh, Jennifer Widom).