Crowdsourcing Effectiveness

Lecture 4

Gianluca Demartini
University of Sheffield

Outline

- Push Crowdsourcing
- Malicious behaviors

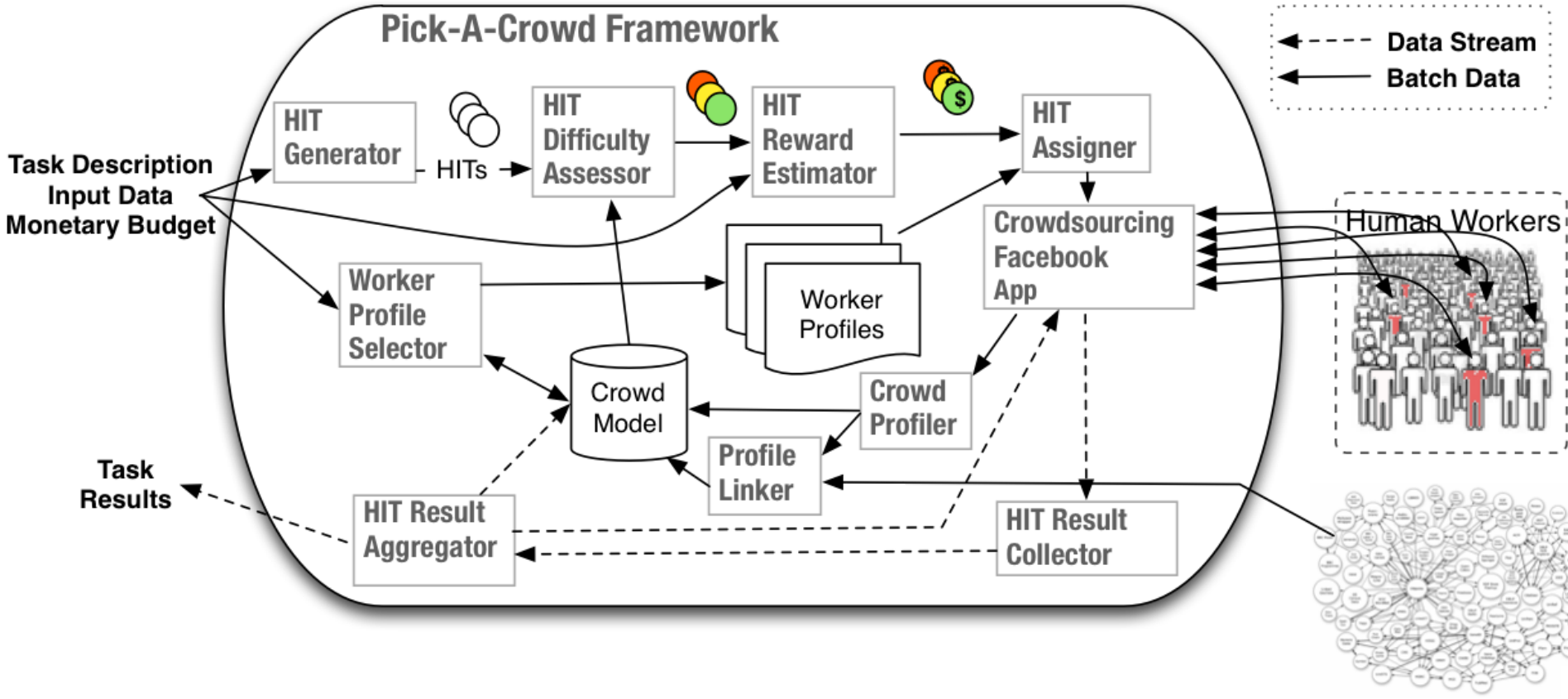
Pull (Traditional) Crowdsourcing

- In MTurk HITs are published on the market
- The first worker willing to do it can take it
- Pro: Fast
- Con: Not necessarily optimal / not the best worker for the task

Push Crowdsourcing

- Pick-A-Crowd: A system architecture that uses Task-*to*-Worker matching:
  - The worker's social profile
  - The task context
- Workers can provide higher quality answers on tasks they relate to

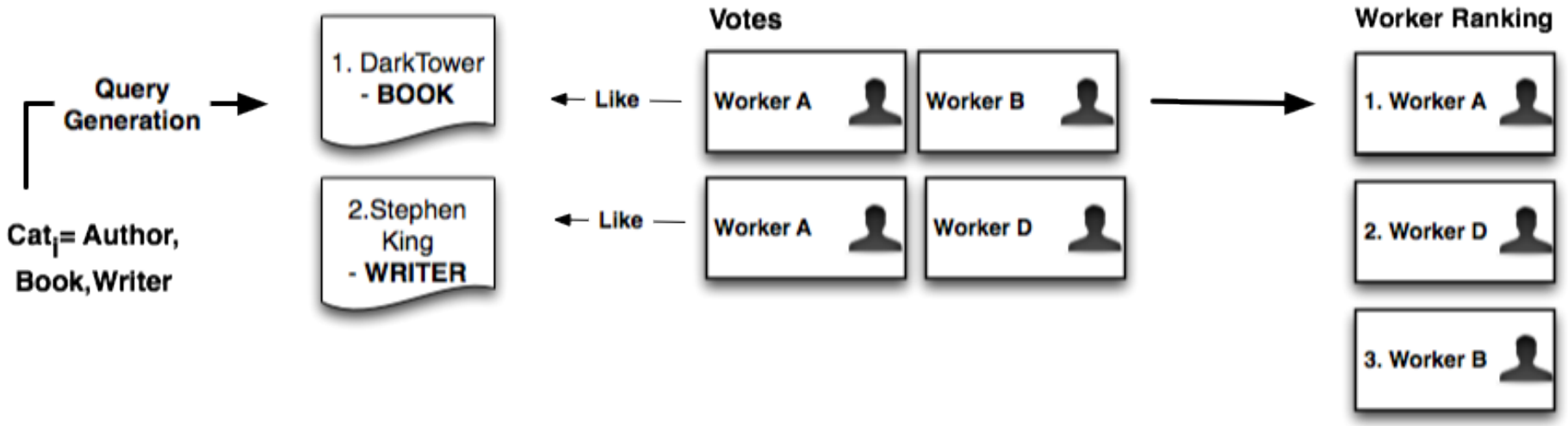
Pick-A-Crowd



Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux.
Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do.
In: 22nd International Conference on World Wide Web (WWW 2013)

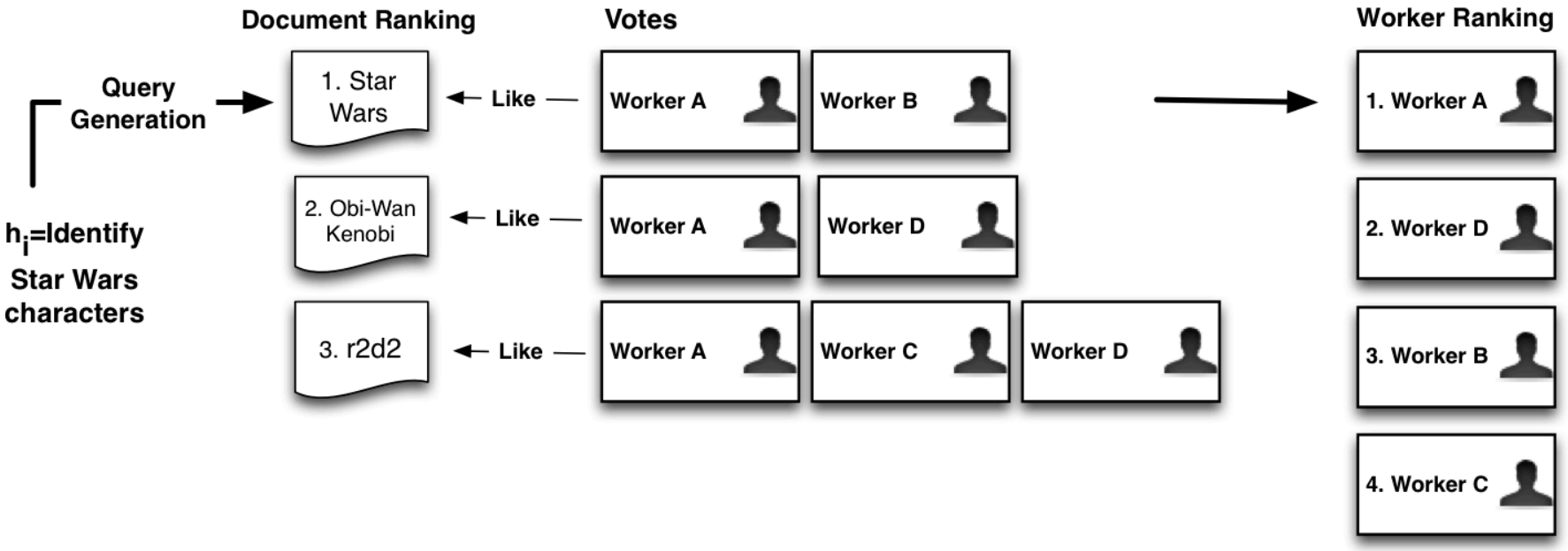
Matching Models (1/3)– Category Based

- The requester provides a list of categories related to the batch
- We create a subset of pages whose category is in the category list of the batch
- Rank the workers by the number of liked pages in the subset



Matching Models– Expert Finding

- Build an inverted index on the *pages*' titles and description
- Use the title/description of the tasks as a key word query on the inverted index and get a subset of *pages*
- Rank the workers by the number of liked pages in the subset



Matching Models (3/3) – Semantic Based

- Link the context to an external knowledge base (e.g., DBPedia)
- Exploit the underlying graph structure to determine the Hits and Pages similarity
  - Assumption that a worker who likes a page is able to answer questions about related entities
  - Worker who likes a page is able to answer questions about entities of the same type
- Rank the workers by the number of liked pages in the subset

Similarity

Relatedness

```
SELECT ?x
WHERE { <uri(a_i)> ?x <uri(p_i)> }.
```

Type-Similarity

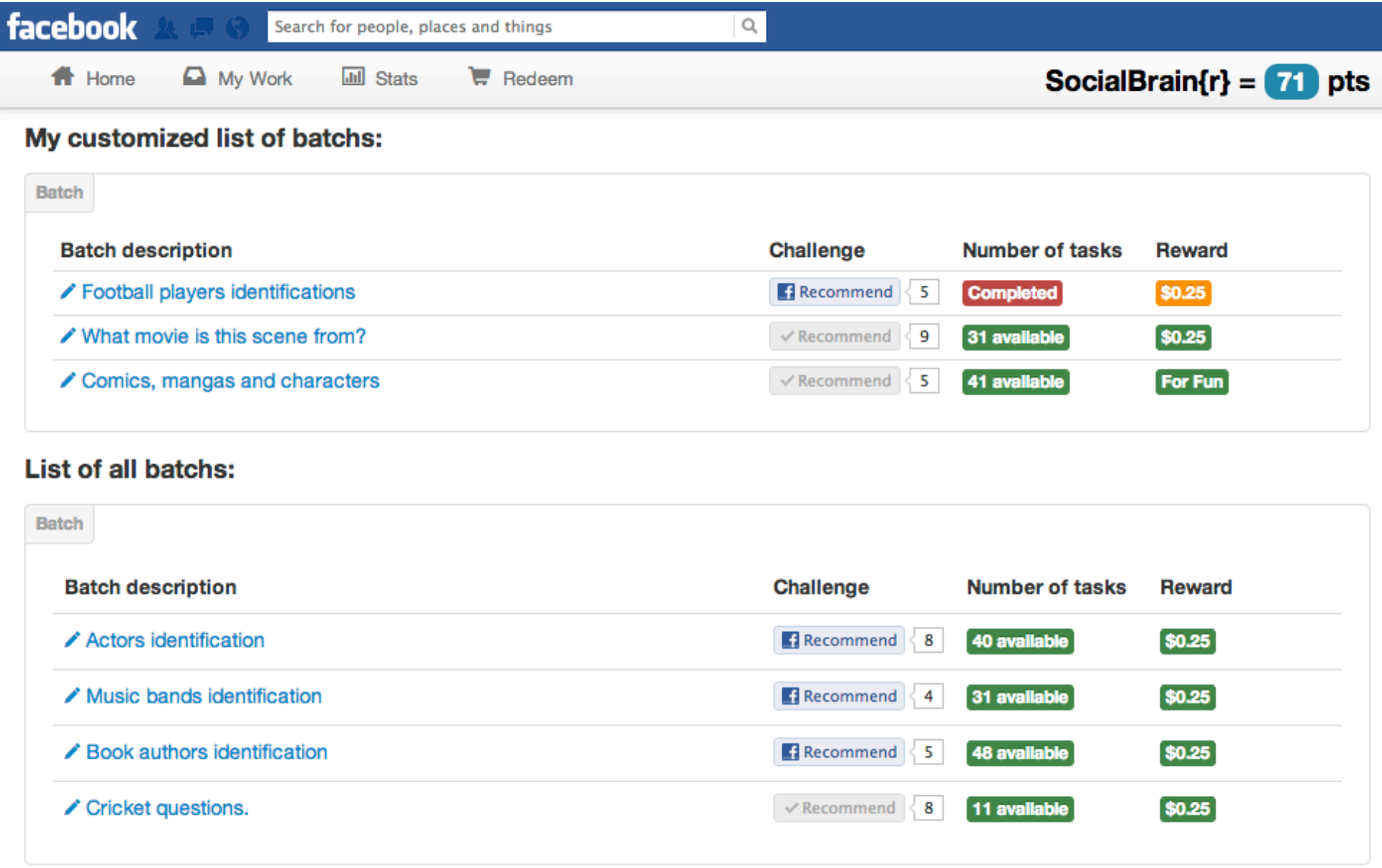
```
SELECT ?x
WHERE { <uri(a_i)> <rdf:type> ?x .
        <uri(p_i)> <rdf:type> ?x
      }.
```

HIT

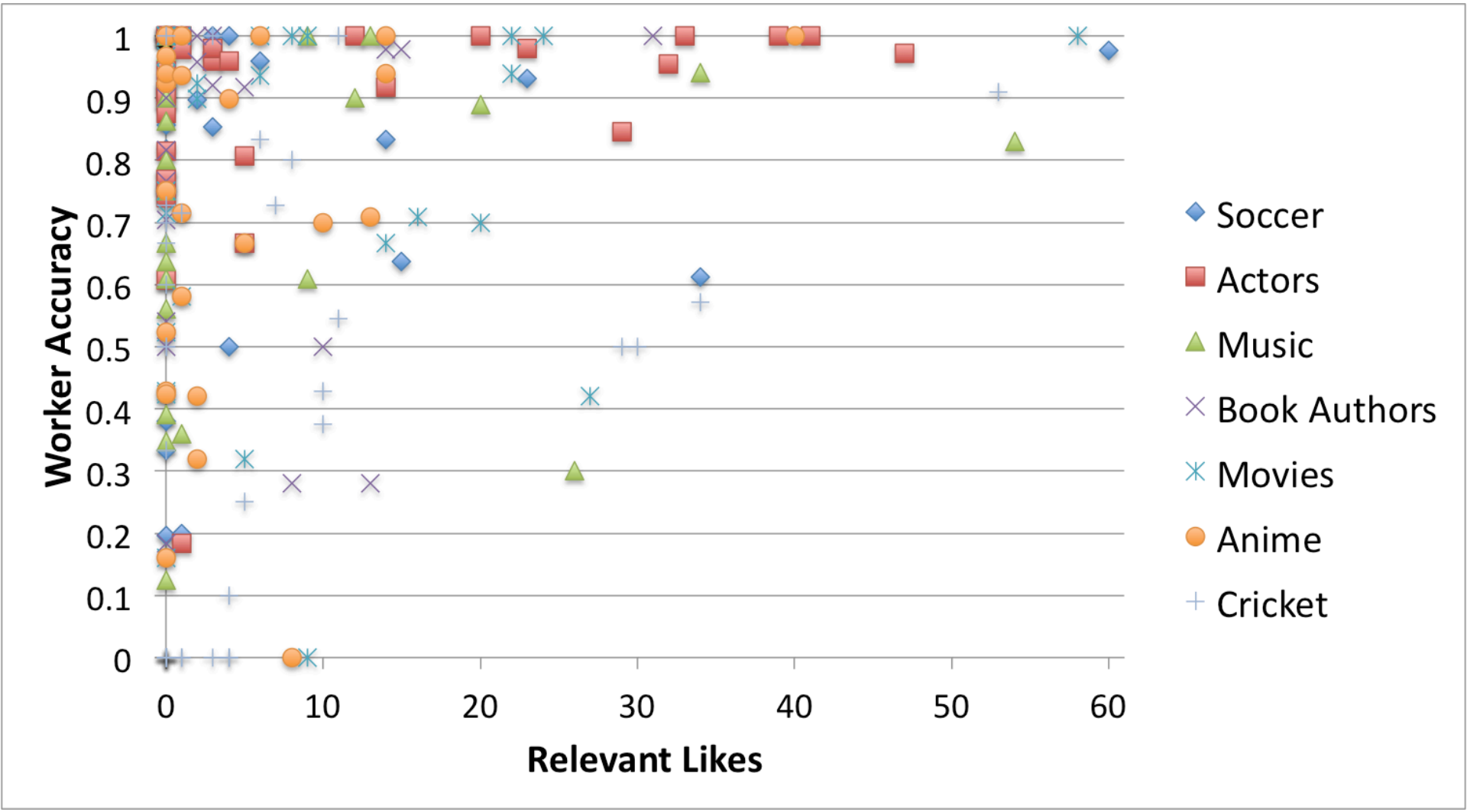
FB Pages

Experimental Evaluation

- The Facebook app **OpenTurk** implements part of the Pick-A-Crowd architecture:
  - More than **170 registered workers** participated
  - Over **12k pages** crawled
- Covered both multiple answer questions as well as open-ended questions
  - 50 images with multiple choice question and 5 candidate answers (Soccer, Actors, Music, Authors,Movies, Animes)
  - Answer 20 open-ended questions related to the topic (Cricket)


facebook
Search for people, places and things
Home
My Work
Stats
Redeem
SocialBrain{r} = 71 pts
My customized list of batchs:
Batch
Batch description
Challenge
Number of tasks
Reward
Football players identifications
Recommend 5
Completed
$0.25
What movie is this scene from?
Recommend 9
31 available
$0.25
Comics, mangas and characters
Recommend 5
41 available
For Fun
List of all batchs:
Batch
Batch description
Challenge
Number of tasks
Reward
Actors identification
Recommend 8
40 available
$0.25
Music bands identification
Recommend 4
31 available
$0.25
Book authors identification
Recommend 5
48 available
$0.25
Cricket questions.
Recommend 8
11 available
$0.25

Like vs Accuracy



Evaluation - Comparison With Mechanical Turk

	Assignment Method	Average Accuracy
AMT	AMT 3	0.66
	AMT 5	0.62
	AMT Masters 3	0.54
PICK-A-CROWD	Category-based 3	0.79
	Category-based 5	0.83
	Voting Model t_i 3	0.80
	Voting Model t_i 5	**0.85**
	Voting Model A_i 3	0.69
	Voting Model A_i 5	0.72
	En. type 3	0.66
	En. type 5	0.79
	1-step 3	0.66
	1-step 5	0.71

Discussion

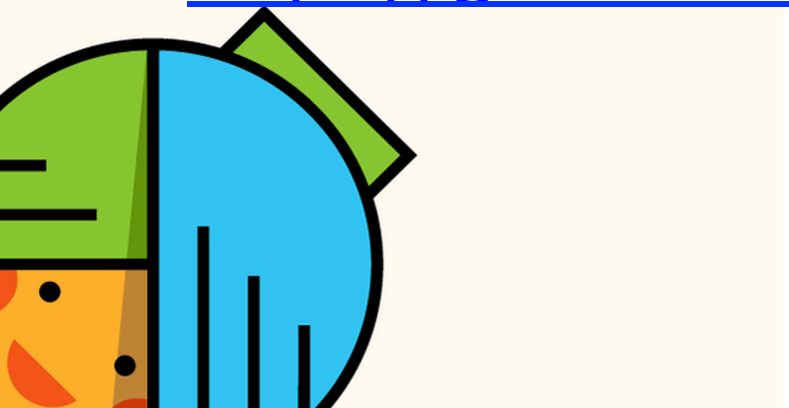
- Pull vs. Push methodologies in Crowdsourcing
- Pick-A-Crowd system architecture with Task-*to*-Worker recommendation
- Experimental comparison with AMT shows a consistent quality improvement

"Workers *Know* what they *Like*"

www.openturk.com

OpenTurk

- Yet another a platform? Build on top of Mturk!
- Chrome Extension for push / notification
- 400+ users
- http://bit.ly/openturk-extension
- Open source: https://github.com/openturk/extension



Understanding Malicious Behaviour in Crowdsourcing Platforms

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. **Understanding Malicious Behaviour in Crowdsourcing Platforms: The Case of Online Surveys**. In: Proceedings of the ACM Special Interest Group on Computer Human Interaction (CHI 2015). Seoul, South Korea, April 2015.

Challenges

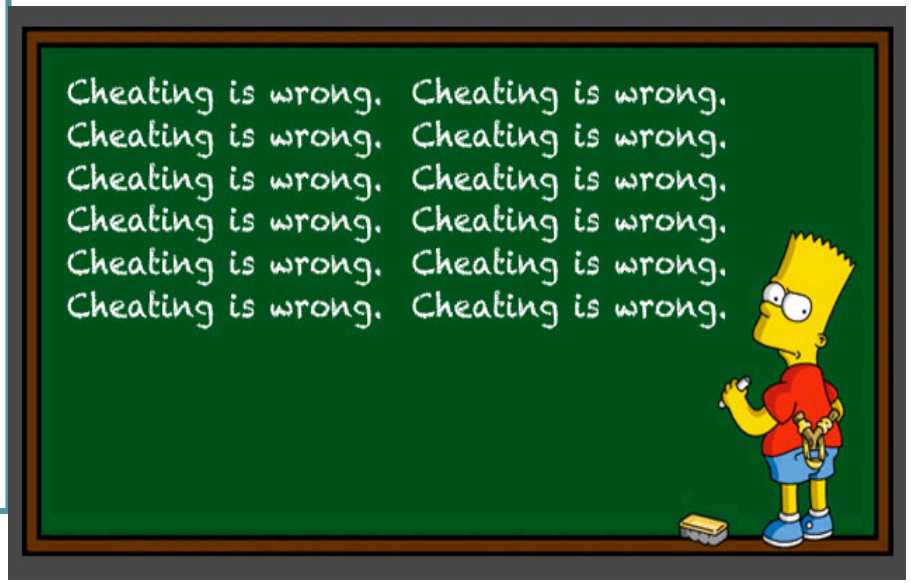
- Quality Control Mechanisms
  - Diverse pool of crowd workers
  - Wide range of behavior
  - Various motivations

Malicious Workers

"workers with ulterior motives, who either simply sabotage a task, or provide poor responses in an attempt to quickly attain task completion for monetary gains"

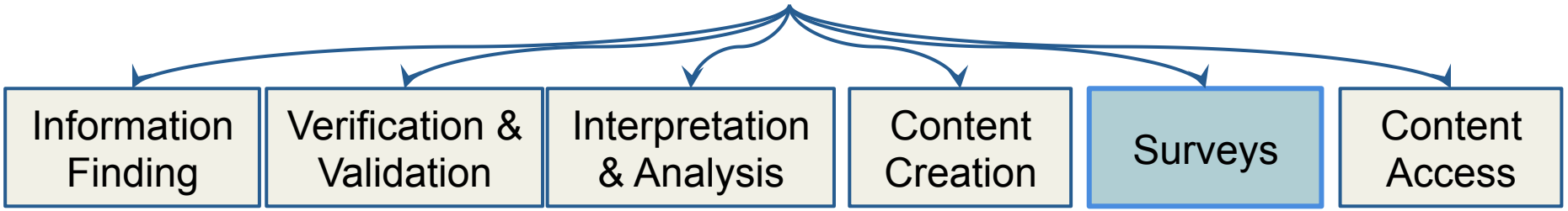
Need to understand workers behavior and types of malicious activity.

- Typically adopted solution to prevent/flag malicious activity : Gold-Standard Questions
- Flourishing Crowdsourcing markets, advances in malicious activity



Background

Taxonomy of Microtasks



- We focus on analyzing the malicious behavior of workers in SURVEYS
  - Subjective nature
  - Open-ended questions
  - Gold-standards are not easily applicable

A Taxonomy of Microtasks on the Web.
Ujwal Gadiraju, Ricardo Kawase and Stefan Dietze. *In Proceedings of the 25th ACM Conference on Hypertext and Social Media. 2014.*

Research Questions

RQ#1 Do untrustworthy workers adopt different methods to complete tasks, and exhibit different kinds of behavior?

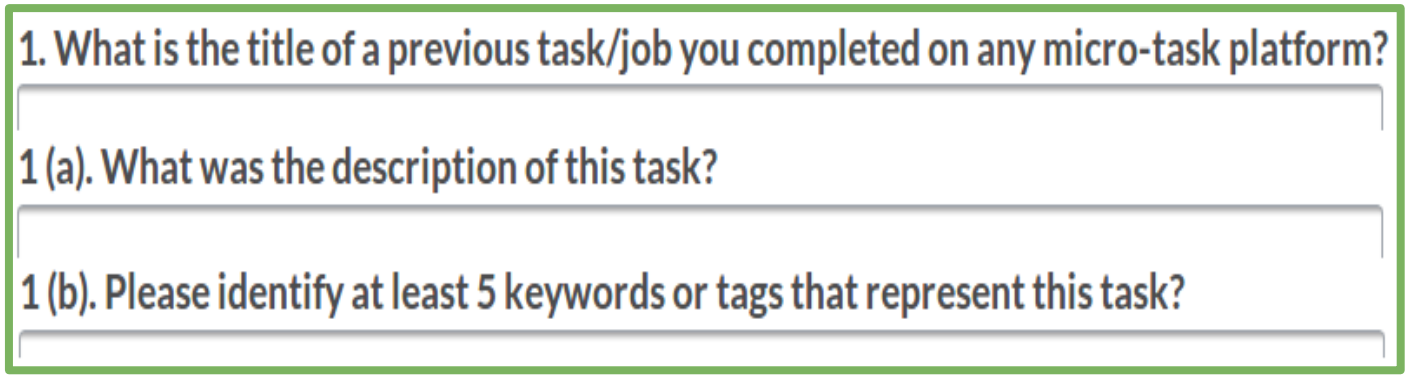
RQ#2 Can behavioral patterns of malicious workers in the crowd be identified and quantified?

RQ#3 How can task administrators benefit from the prior knowledge of plausible worker behavior?

Survey Design

- CrowdFlower Platform to deploy survey
- Survey questions
  - Demographics
  - Educational & general background
- 34 Questions in total
  - Open-ended
  - Multiple Choice
  - Likert-type
- Responses from 1000 crowd workers
  - Monetary Compensation per worker : 0.2 USD

- Questions regarding previous tasks that were successfully completed



- 2 Attention-check questions
  - Engage workers
  - Gold-standard to separate Trustworthy/Untrustworthy workers (we found **568** trustworthy, **432** untrustworthy)

Analyzing Malicious behavior in the Crowd

Based on the following aspects, we investigate the behavioral patterns of crowd workers.

I. eligibility of a worker to participate in a task
II. conformation to the pre-set rules
III. satisfying expected requirements fully

Behavioral Patterns

Ineligible Workers (IW)

Instruction: Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.

Response: *'this is my first task'*

Fast Deceivers (FD)

eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.

Response: *'What's your task?'* , *'adasd'*, *'fgfgf gsd ljlkj'*

Rule Breakers (RB)

Instruction: Identify 5 keywords that represent this task (separated by commas).

Response: *'survey, tasks, history'* , *'previous task yellow'*

Smart Deceivers (SD)

Instruction: Identify 5 keywords that represent this task (separated by commas).

Response: *'one, two, three, four, five'*

Gold Standard Preys (GSP)

These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions!

Observations

We manually annotated each response from the 1000 workers.

- ➢568 workers passed the gold-standard: Trustworthy workers (TW)
- ➢432 workers failed to pass the gold-standard: Untrustworthy workers (UW)
- ➢335 trustworthy workers gave perfect responses: Elite workers
- ➢665 non-elite workers (233 TW, 432 UT) were manually classified into the different classes according to their behavioral patterns.

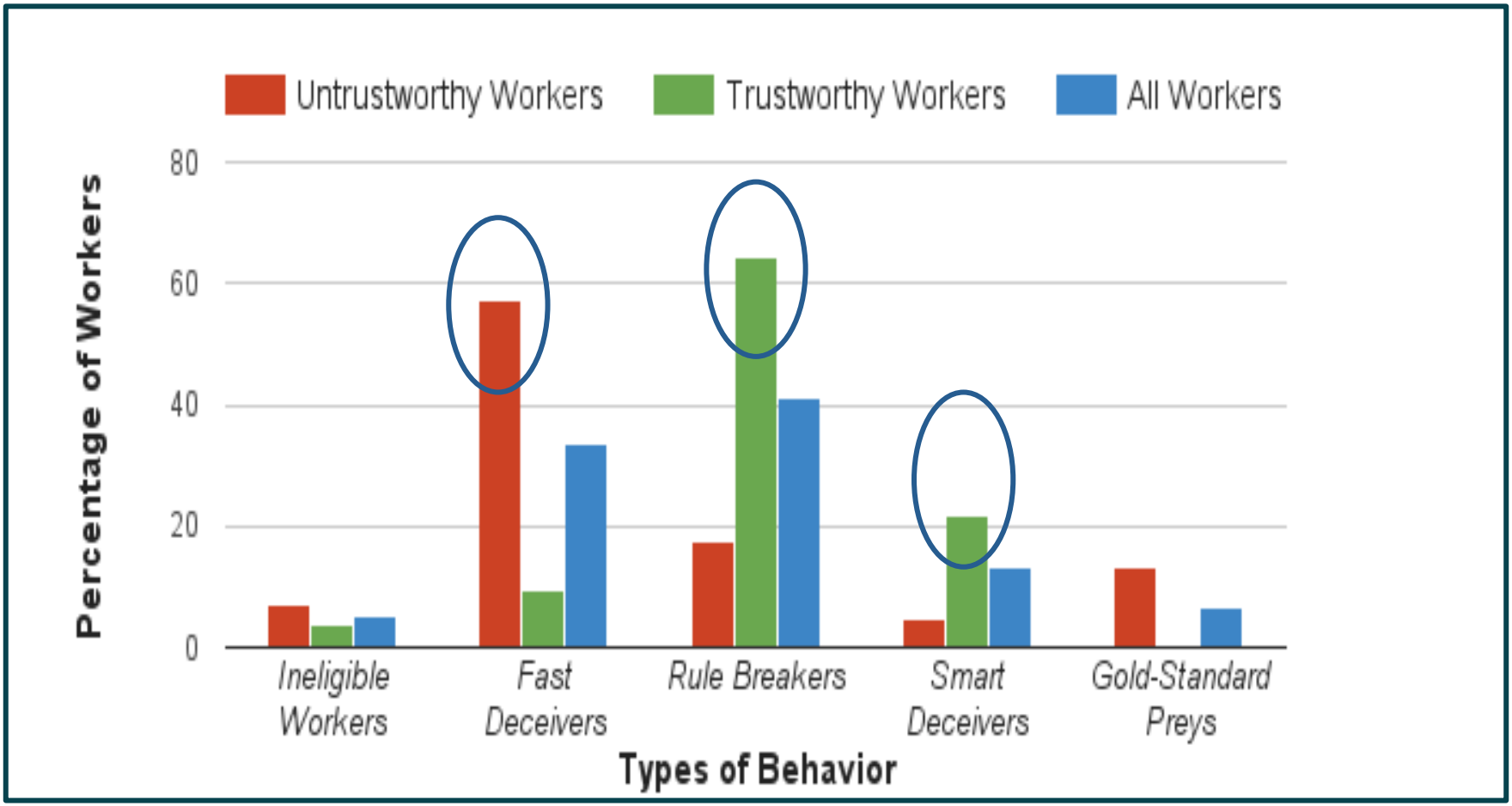
Workers Classification

- 73 untrustworthy workers and 93 trustworthy workers were classified into 2 different classes, while the rest were uniquely classified.
- Inter-rater agreement between the experts (according to Krippendorf's Alpha) : 0.94

Acceptability of Responses

- Inter-rater agreement between the experts (according to Krippendorf's Alpha) : 0.89

Distribution of Workers



Measuring the Maliciousness of workers

Acceptability: "The acceptability of a response can be assessed based on the extent to which a response meets the priorly stated expectations."
E.g.

Instruction: Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.

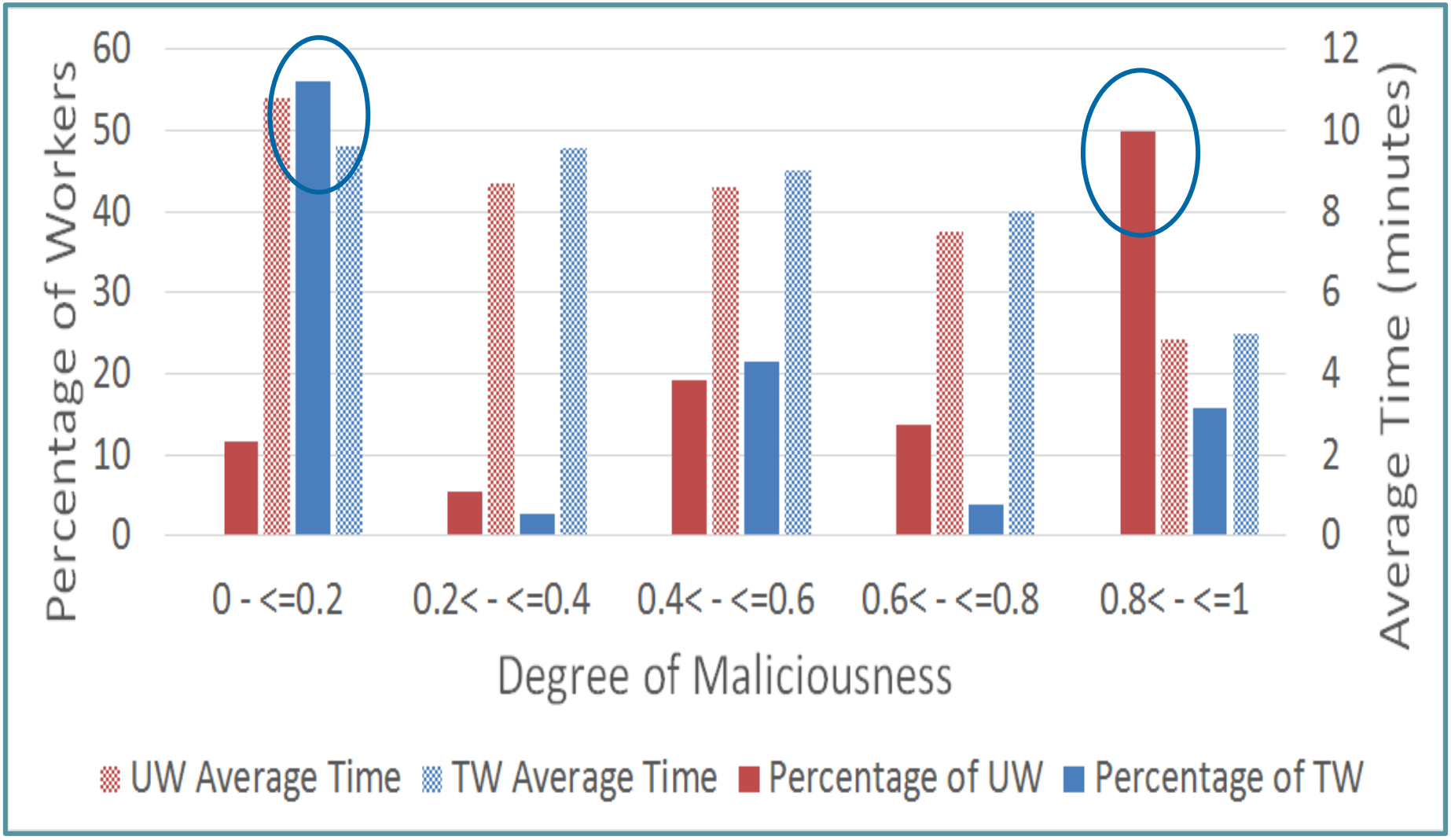
Response: '*survey, tasks, history*' ⇒ '0'

Response: '*previous, job, finding, authors, books*' ⇒ '1

We consider open-ended questions.

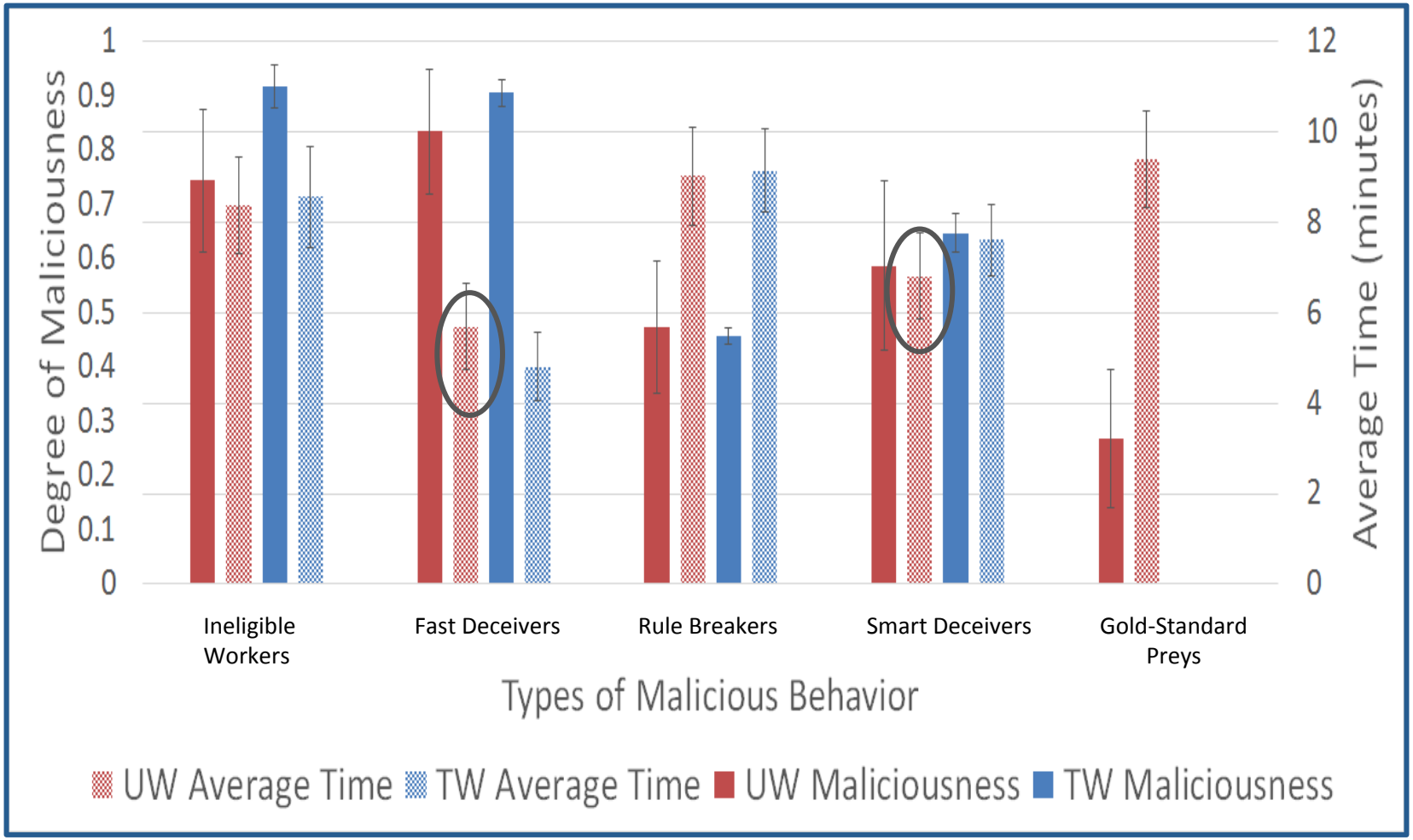
$$M_{worker} = 1 - (1/n \sum_{i=1}^{n} A_{r_i})$$

where, *n* is the total number of responses from a worker and A_{ri} represents the acceptability of response 'i'

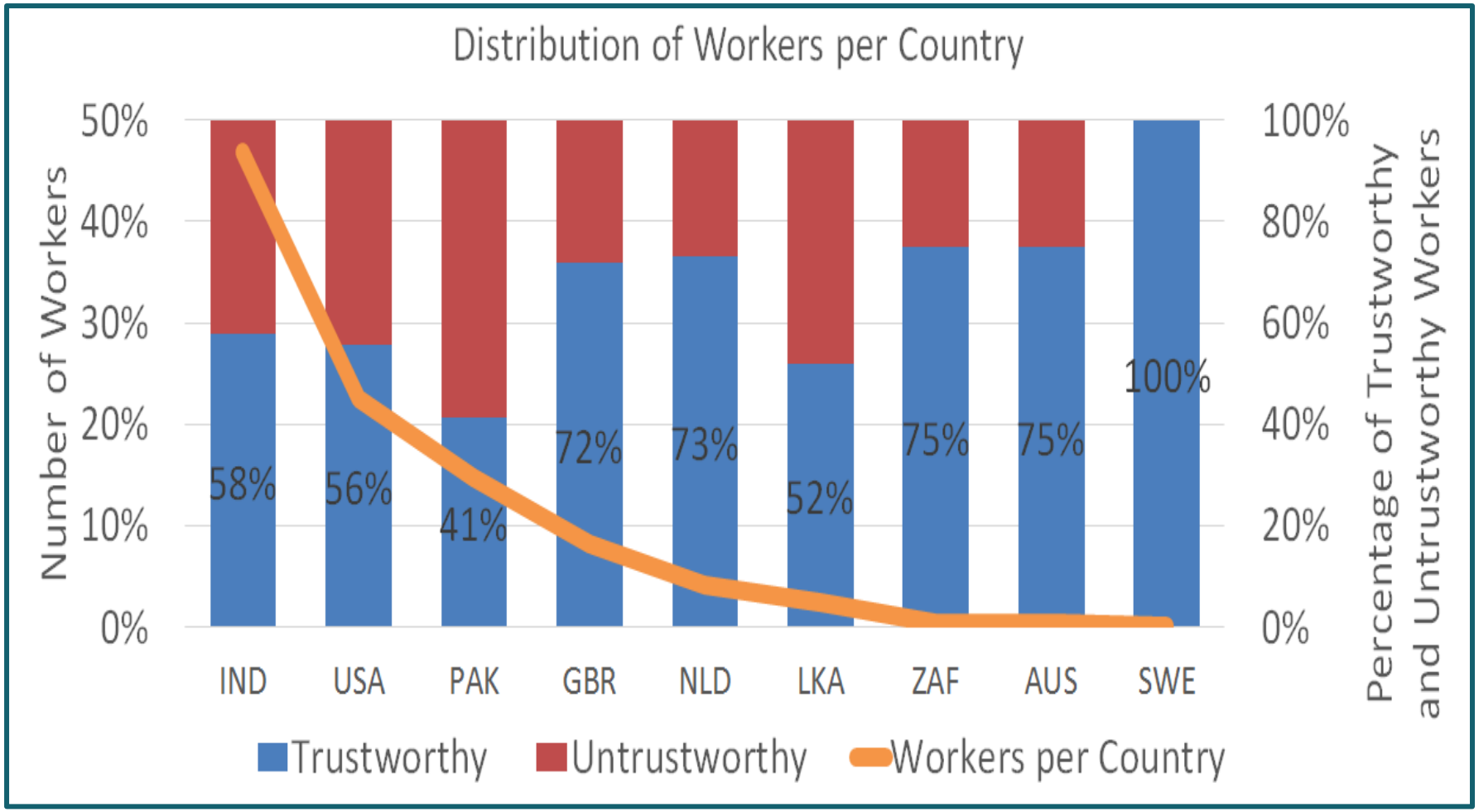


Degree of maliciousness of trustworthy (TW) and untrustworthy workers (UW) and their average task completion time.

Task Completion Time vs Worker Maliciousness



Where are the workers from?



Tipping Point

"the first point at which a worker begins to exhibit malicious behavior after having provided an acceptable response"

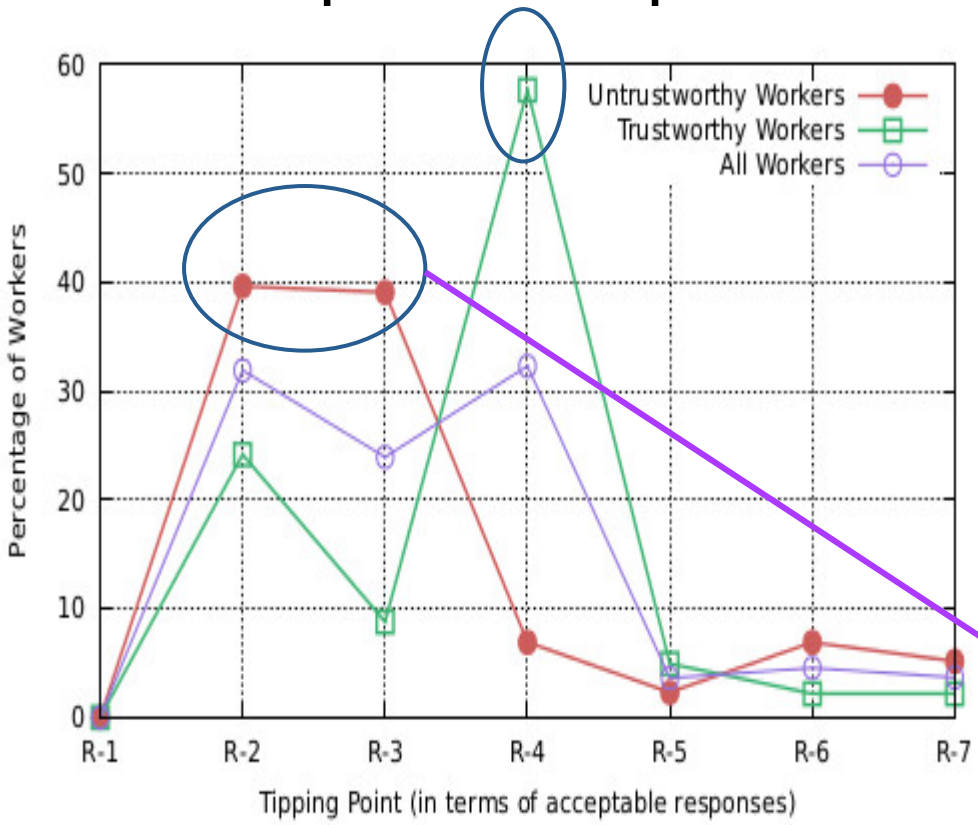


Table 1. Relationship between the Maliciousness and Tipping Point of untrustworthy and trustworthy workers (percentage of workers having tipping point @R).

Maliciousness	UW	TW
$0 < M \leq 0.2$	40.9% @ R-7 31.8% @ R-6	28.5% @ R-7 28.5% @ R-5
$0.2 < M \leq 0.4$	43.47% @ R-3 21.73% @ R-6	30% @ R-5 30% @ R-3
$0.4 < M \leq 0.6$	66.19% @ R-3 25.35% @ R-2	88% @ R-4 5.1% @ R-3
$0.6 < M \leq 0.8$	71.05% @ R-2 28.95% @ R-3	60% @ R-3 40% @ R-2
$0.8 < M \leq 1$	100% @ R-2	100% @ R-2

Task Design Guidelines

- Using the 'Tipping Point' for early detection of malicious activity.
- Using 'Malicious Intent' as a measure to discard unreliable responses from workers and improve the quality of results.

Ineligible Workers

Pre-screening to tackle Ineligible Workers (IW).

Fast Deceivers

Rule Breakers

Stringent and persistent validators and monitoring worker progress to tackle Fast Deceivers (FD) and Rule Breakers (RB).

Gold Standard Preys

Smart Deceivers

Psychometric approaches to tackle Smart Deceivers (SD).

Post-processing to accommodate fair responses from Gold-standard Preys (GSP).

Contributions

- ✓ Identified different types of malicious behavior exhibited by crowd workers. — RQ#1
- ✓ Measuring ‘maliciousness’ of workers to quantify their behavioral traits, and ‘tipping point’ to further understand worker behavior. — RQ#2
- ✓ This understanding helps requesters in effective task design, ensures adequate utilization of the crowdsourcing platform(s).
- ✓ Guidelines for effective design of Surveys by limiting malicious activity. — RQ#3

Summary

- Design the User Interfaces
- Define the right incentives
- Task Patterns
- Quality control
  - Task design
  - Mechanisms: honeypots, agreement, redundancy, answer aggregation, pricing
  - Know the crowd
  - Understand human behavior