Task Assignment Optimization in Crowdsourcing
(and its applications to crisis management)

Sihem Amer-Yahia
DR CNRS @ LIG
Sihem.Amer-Yahia@imag.fr
joint work with:
Soudip Roy Chowdhury (INRIA), Carlos Castillo, Muhammad Imran (QCRI), Muhammad Rizwan Asghar (Saarland U.), Senjuti Basu Roy (UW Tacoma), Gautam Das, Habibur Rahman, Saravanan Thirumuruganathan (UT Arlington)

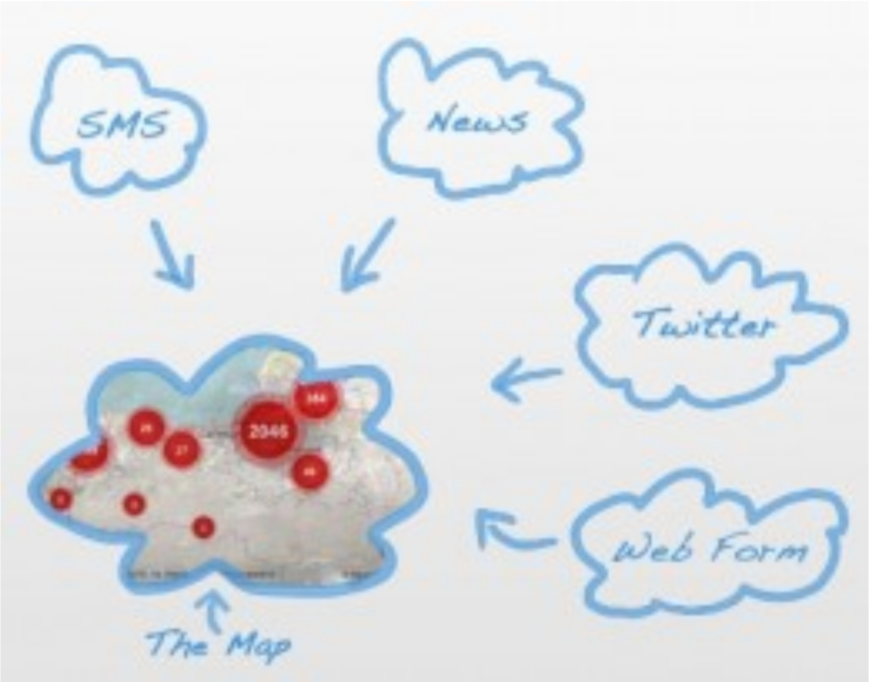
SAGEO conference
Grenoble Nov 26th, 2014

	Traditional Media	**Today's Reporting**	
Crisis awareness and reporting	Onsite reporters	Tweets, text messages, news reports	
Report timeliness	Mixed (reliance on experts; limited resources)	✔	
Report quality	✔	Mixed (noisy; rumor and misinformation)	
Report cost	High	✔	

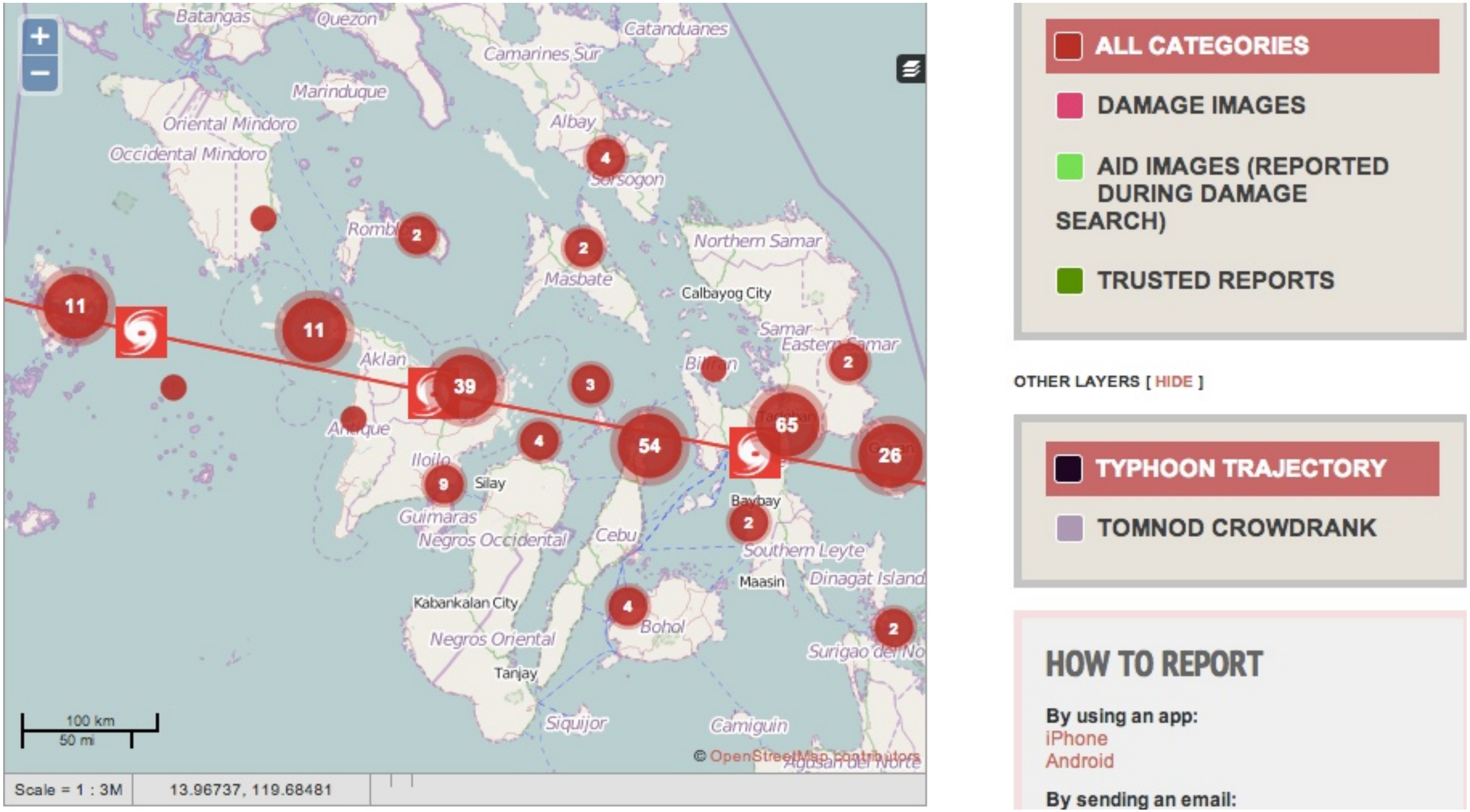
	Traditional Media	**Today's Reporting**	**Future Reporting?**
Crisis awareness and reporting	Onsite reporters	Tweets, text messages, news reports	Online tweet analysis discovers and follows course of incident
Report timeliness	Mixed (reliance on experts; limited resources)	✔	✔ (if detection of crisis is done on time)
Report quality	✔	Mixed (noisy; rumor and misinformation)	✔ (via careful assignment)
Report cost	High	✔	✔ (within budget)

CrowdMap from Ushahidi

- ***Ushahidi*** *is a non-profit software company that develops free software for information collection, visualization, and interactive mapping via citizen journalism.*
- *Started with Kenya's disputed 2007 presidential election that collected eyewitness reports of violence reported by email and text message and placed them on a map.*
- ***CrowdMap****: a place where volunteers go to "check-in" and add data about incident.*
- ***Future option for CrowdMap****: start with a world map that shows crisis hotspots (determined by geo-tagged data).*



CrowdMap from Ushahidi



Talk outline

1. Quick overview of existing crowdsourcing
2. Task assignment
3. Online tweet monitoring

Crowdsourcing

- **Crowdsourcing: a variety of tasks**
  - *Micro-tasks*: data gathering (e.g. picture/video tagging, opinion solicitation (e.g. restaurant ratings,)
  - *Collaborative tasks*: document editing (e.g., Wikipedia), creative design, fansubbing, solution outsourcing (e.g., Netflix contest)
- **Existing systems**
  - *Platforms*: AMT, Turkit, Innocentive, CloudFlower, etc.
  - *Crowd*: volatile, asynchronous arrival/departure, various levels of attention/accuracy/expertise
- **3 primary processes**
  - Worker skill estimation
  - Worker-to-task assignment
  - Task accuracy evaluation

Challenges

- **Who Evaluates What and How?**
- **How to Estimate Worker Skills?**
- **How to Assign Tasks to Workers?**
- **How to do all of the above efficiently?**

- **Magnified by:**
  - Human factors
  - Scale

Limitations

- **Developing a crisis-specific platform for each disaster situation is costly**
- **Recent research undertakes some challenges in silo, for specific cases: e.g. real-time crowdsourcing, highly volatile crowds, single worker skill**
  - Active learning strategies for task accuracy improvement [Boim et. Al. 2012, Krager et. al. 2011, Ramesh et. al. 2012]
  - Worker-to-task-assignment [Ho et. al. 2012]
- **Human involvement introduces uncertainty**
  - *Worker availability*
  - *Worker wage*: deviations even among persons of the same profile, due to workload, time
  - *Worker skill*: may decline with workload, change with motivation

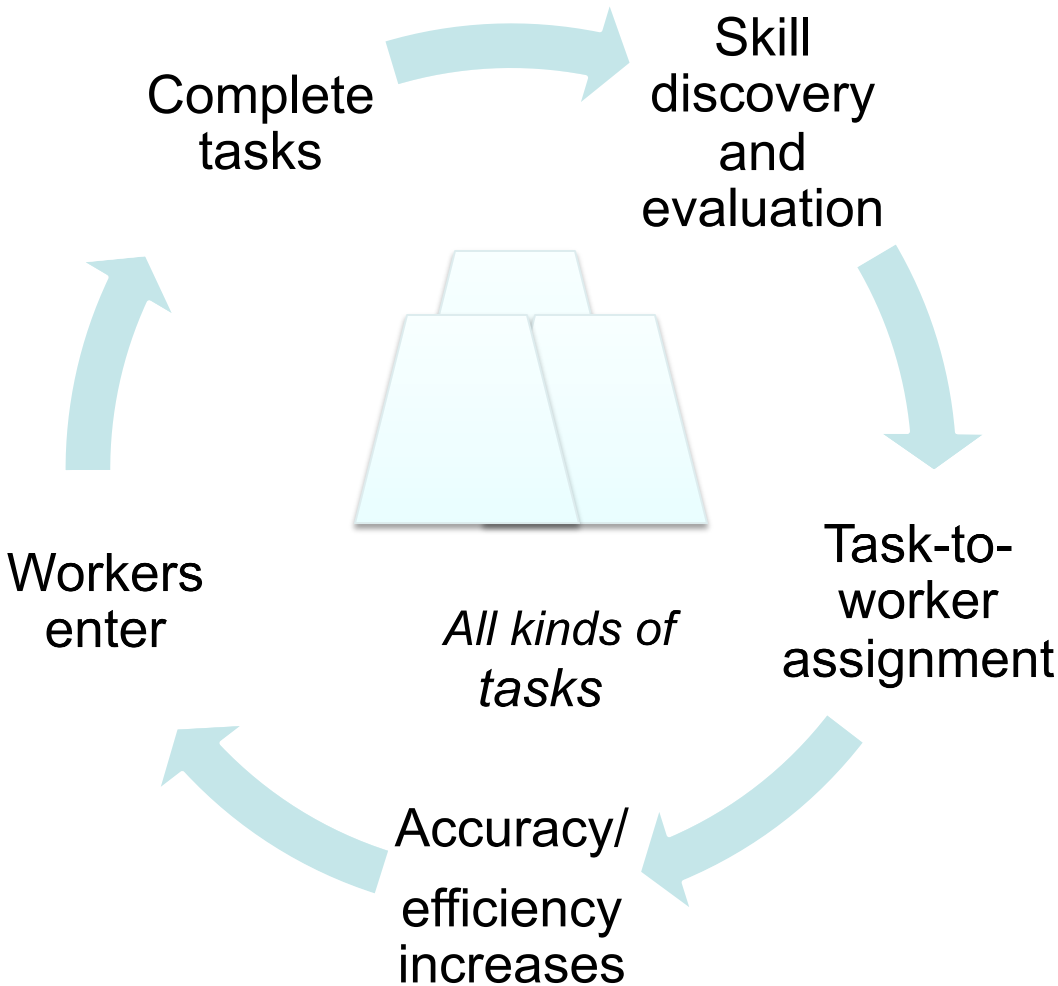
Talk outline

1. Quick overview of existing crowdsourcing
2. Task assignment
3. Online tweet monitoring

Optimization in Knowledge-Intensive Crowdsourcing

@VLDBJ 2015 (to appear)

Integrated Architecture



Expressing Task Assignment

- **Input: tasks to complete, human workers**
- **Output: completed tasks**

- **Each task has skill/quality/cost requirements**
- **Each worker has human factors: *skill, expected wage, acceptance ratio***

- **Desirable properties:**
  - ***Task-centric***: high quality tasks (relevant workers), low cost
  - ***Worker-centric***: balanced workload, good incentive (high pay, relevant tasks)
  - ***System-centric***: low latency

Example: Maximize task quality under task-centric and worker-centric constraints

objective: maximize aggregated v_t

$$\text{Maximize } \mathcal{V} = \Sigma_{\forall t \in T} v_t$$

aggregated worker skills and wages

task quality constraint

$$v_t = \begin{cases} W_1 \times \Sigma_{\forall j \in \{1..m\}} q_{t_j} + W_2 \times (1 - \frac{w_t}{W_t}) & \text{if } q_{t_j} \geq Q_{t_j} \\ & \wedge\ w_t \leq W_t \\ 0 & \text{if } q_{t_j} < Q_{t_j} \\ & \vee\ w_t > W_t \end{cases}$$

where $W_1, W_2 \geq 0$ and $W_1 + W_2 = 1$.

task budget

Task Assignment Solution Overview

Task Assignment Problem is NP-hard (reduction using Multiple-Knapsack Problem)

Our approach:

- **Offline – Index Building for a workload of tasks**
- **Online – Index Maintenance when tasks occurs**
  - How to replace a worker who is not available or does not accept a task?

Optimal Solution Offline Index Building

IP-based

$$q_{t_j} = \Sigma_{\forall u \in \mathcal{U}} u_t \times p_u \times u_{s_j} \geq Q_{t_j}, \forall j \in \{1..m\}$$

$$w_t = \Sigma_{\forall u \in \mathcal{U}} u_t \times p_u \times w_u \leq W_t$$

$$u_t = [0/1]$$

$$X_l \leq \Sigma_{\forall t \in T} \{u_t\} \leq X_h$$

Approximation Solution for Offline Index Building

- **Objective function submodular and becomes monotonic when W2 = 0**

- **Contribution to index building**
  - A greedy deterministic algorithm with a 1-1/e approximation factor when submodular and monotonic
  - A greedy randomized algorithm with a 2/5 approximation factor when submodular
- **Contribution to index maintenance**
  - Solve a marginal IP
  - Cluster workers to reduce size

Experimental Evaluation

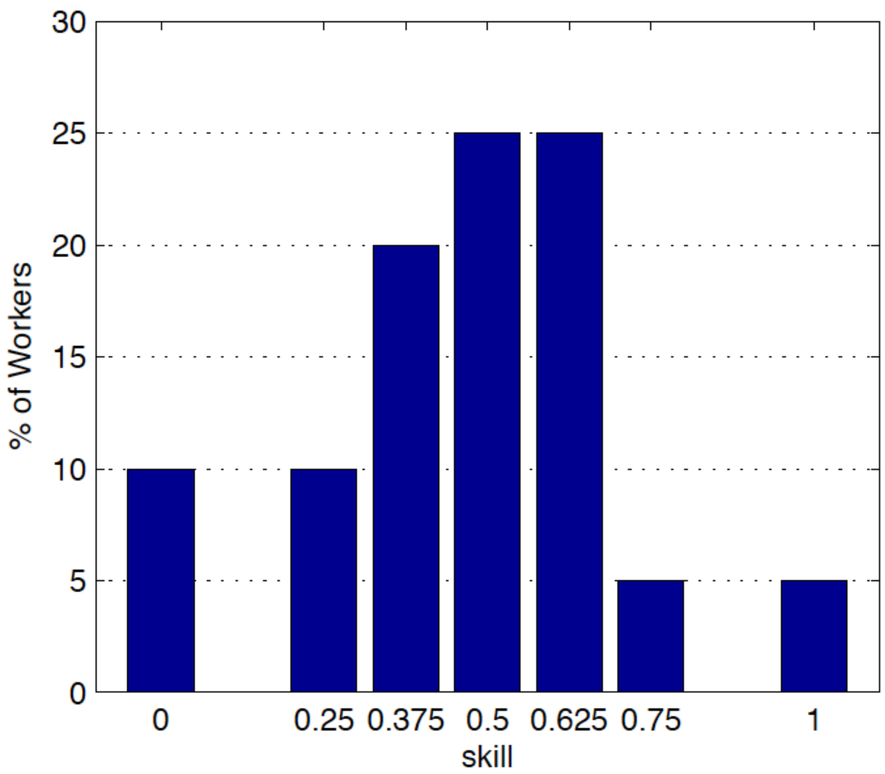
- **Quality Experiments using multiple Applications**
  - Collaborative Document Editing
  - 20 workers asked to produce reports on 5 different topics:
  1) Political unrest in Egypt, 2) NSA document leakage, 3) Playstation games, 4) All electric cars and 5) Global warming

- **Scalability experiments**
  - A collaborative crowd simulator

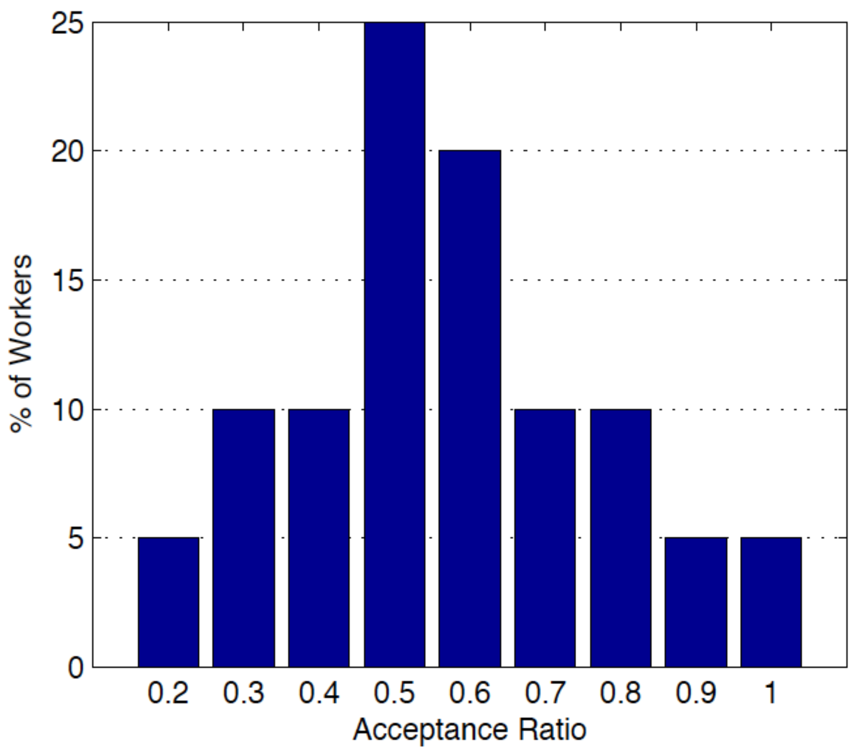
Quality Experiments

- **A total of 230 workers hired on AMT**
- **A set of 8 multiple choice questions per task, to assess skills**
- **Study conducted in multiple phases**
  - Phase1- Skill and Cost of workers learned using benchmark dataset
  - Phase2- Task assignment
  - Phase3- Completed tasks evaluated by crowd workers

AMT worker distributions (Egypt task)

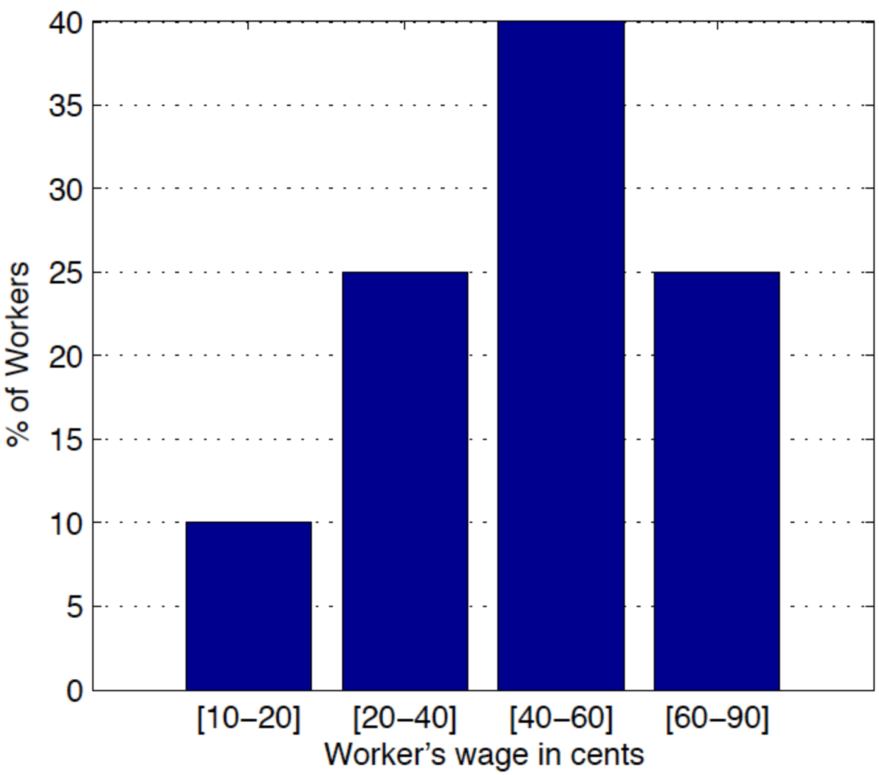


(a) Skill distribution

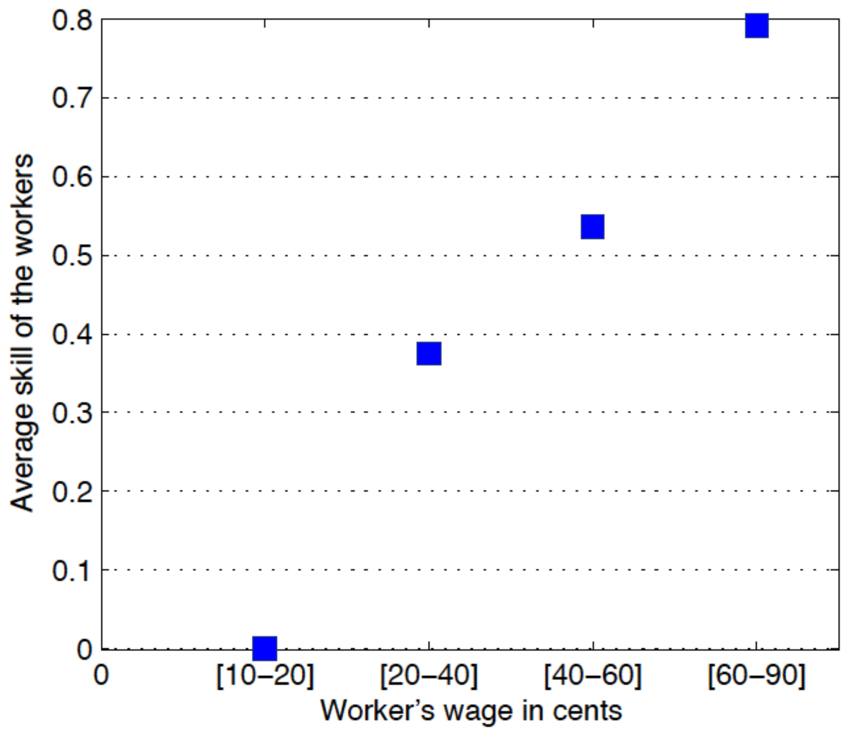


(b) Acceptance ratio distribution

AMT worker distributions (Egypt task)



(c) Wage distribution



(d) Strong positive correlation between worker skill and wage

Quality Assessment

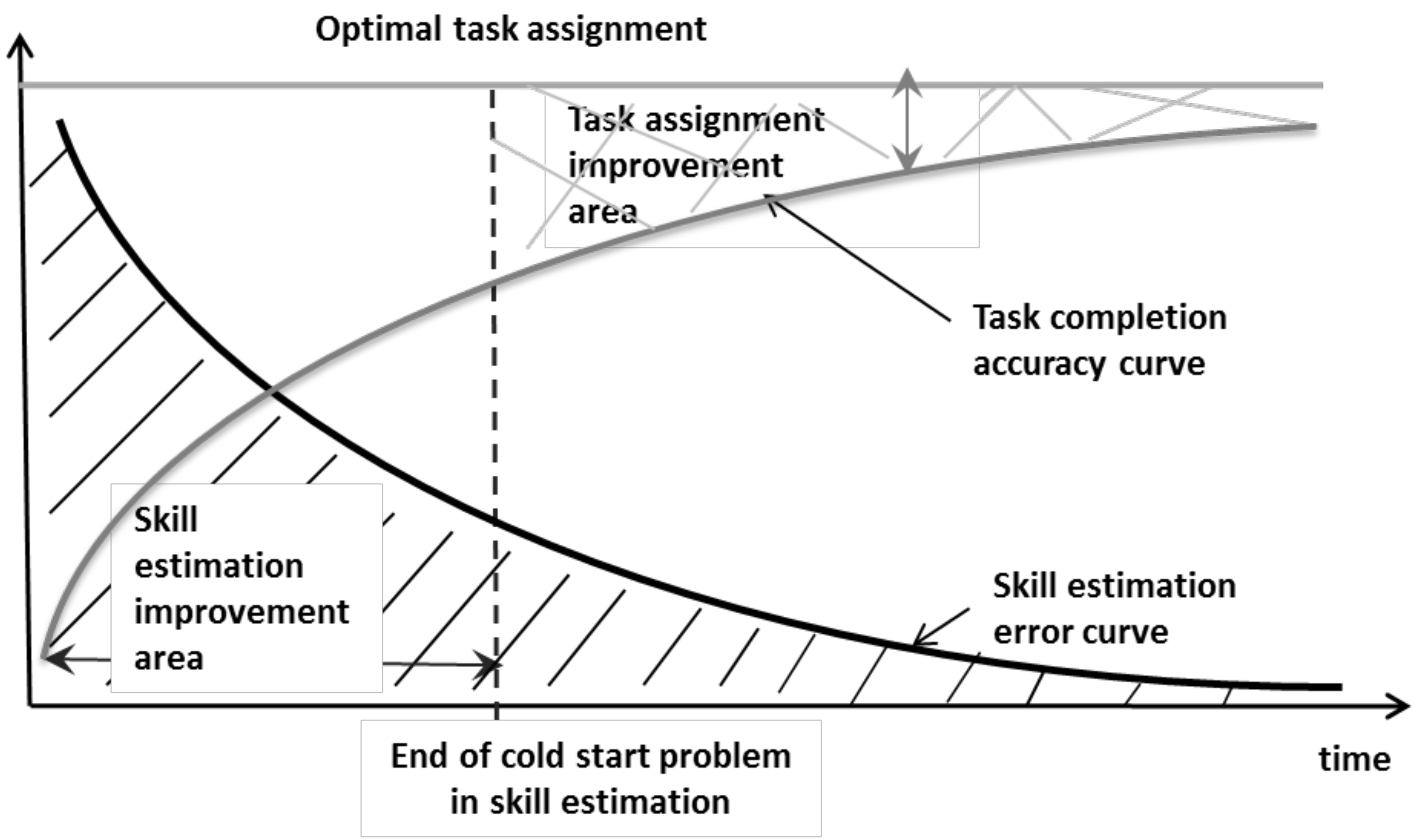
- Scale of 1-5 by 150 AMT workers
- Compared to Benchmark and Online-greedy

Average Rating							
Task	Algorithm	Completeness	Grammar	Neutrality	Clarity	Timeliness	Added-value
Egypt political unrest	SmartCrowd	4.5	4.2	4.0	4.2	4.1	4.0
	Online-greedy	3.3	3.4	3.3	3.0	3.4	3.1
	Benchmark	3.1	3.2	3.3	3.1	3.2	2.9
NSA document leakage	SmartCrowd	4.5	4.7	4.3	3.9	4.1	4.1
	Online-greedy	3.2	3.4	3.3	3.3	3.0	2.9
	Benchmark	3.3	3.3	3.4	2.9	2.9	3.4
PS Games	SmartCrowd	4.3	4.2	4.0	4.1	4.1	4.2
	Online-greedy	3.2	3.3	3.3	3.1	3.0	2.9
	Benchmark	3.0	3.2	3.1	2.8	2.9	2.9
All electric cars	SmartCrowd	4.2	4.2	4.1	4.4	4.0	4.1
	Online-greedy	3.0	3.1	3.3	3.0	2.9	2.8
	Benchmark	2.9	2.6	2.6	3.0	2.8	2.3
Global warming	SmartCrowd	4.2	4.3	4.5	4.2	4.1	3.7
	Online-greedy	3.0	3.2	3.1	3.4	3.3	3.3
	Benchmark	2.9	2.9	3.1	3.2	2.9	2.7

Summary of Quality Experiments (from translation task)

- Higher affinity impacts positively quality
- A large group (beyond size 10) is less effective
- Region-based affinity is more effective than age-gender based

Opportunities



	Traditional media	Basic Crowdsourcing	ECCO
Crisis awareness and reporting	Onsite reporters	Tweets, text messages, news reports	Online tweet analysis discovers and follows course of incident
Task timeliness	Mixed (reliance on experts but limited resources)	✔	✔ (if detection of crisis is done on time)
Task quality	✔	Mixed (noisy; rumor and misinformation)	✔ (via careful assignment)
Task cost	High	✔	✔ (within budget)

Talk outline

1. **Quick overview of existing crowdsourcing**
2. **Task assignment**
   – Flexible optimization framework with task-centric, worker-centric goals and constraints
   – Applicable to collaborative tasks
3. **Online tweet monitoring**

Tweet4act

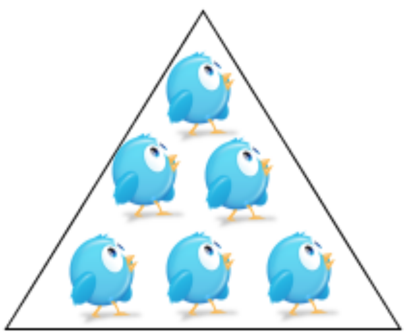
Using Incident-Specific Profiles for Classifying Crisis-Related Messages

@ISCRAM 2013

Classifying incident-related tweets

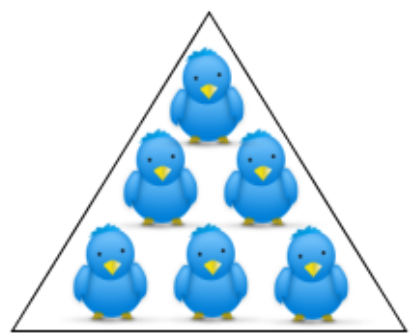
1. Identify messages related to an incident.
2. Classify incident-messages with the corresponding period (PRE, DURING, POST)
3. Apply it to data on the fly

PRE-INCIDENT



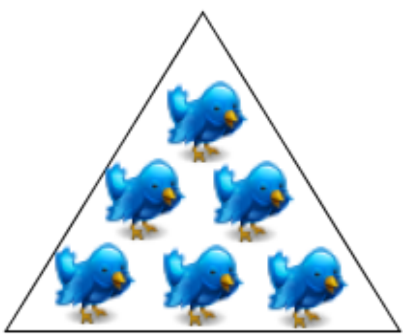
- *Caution, warnings*
- *Alerts etc.*

DURING-INCIDENT



- *Damage*
- *Causalities etc.*

POST-INCIDENT



- *Request for help*
- *Donations etc.*

Disaster-related Tweets

- ***[PRE]*** *New #tropical storm forms in the West #Pacific. #Nesat may hit the #Philippines & #China as a #typhoon next week*
- ***[DURING]*** *@Yahoo News: Powerful #typhoon with winds up to 106 mph makes landfall in #Philippines as 100,000 odered to fless homes*
- ***[POST]*** *News5 Action center is now accepting donations for the victims of Typhoon "pedring. Drop boxes are located @ TV5 Office :)*

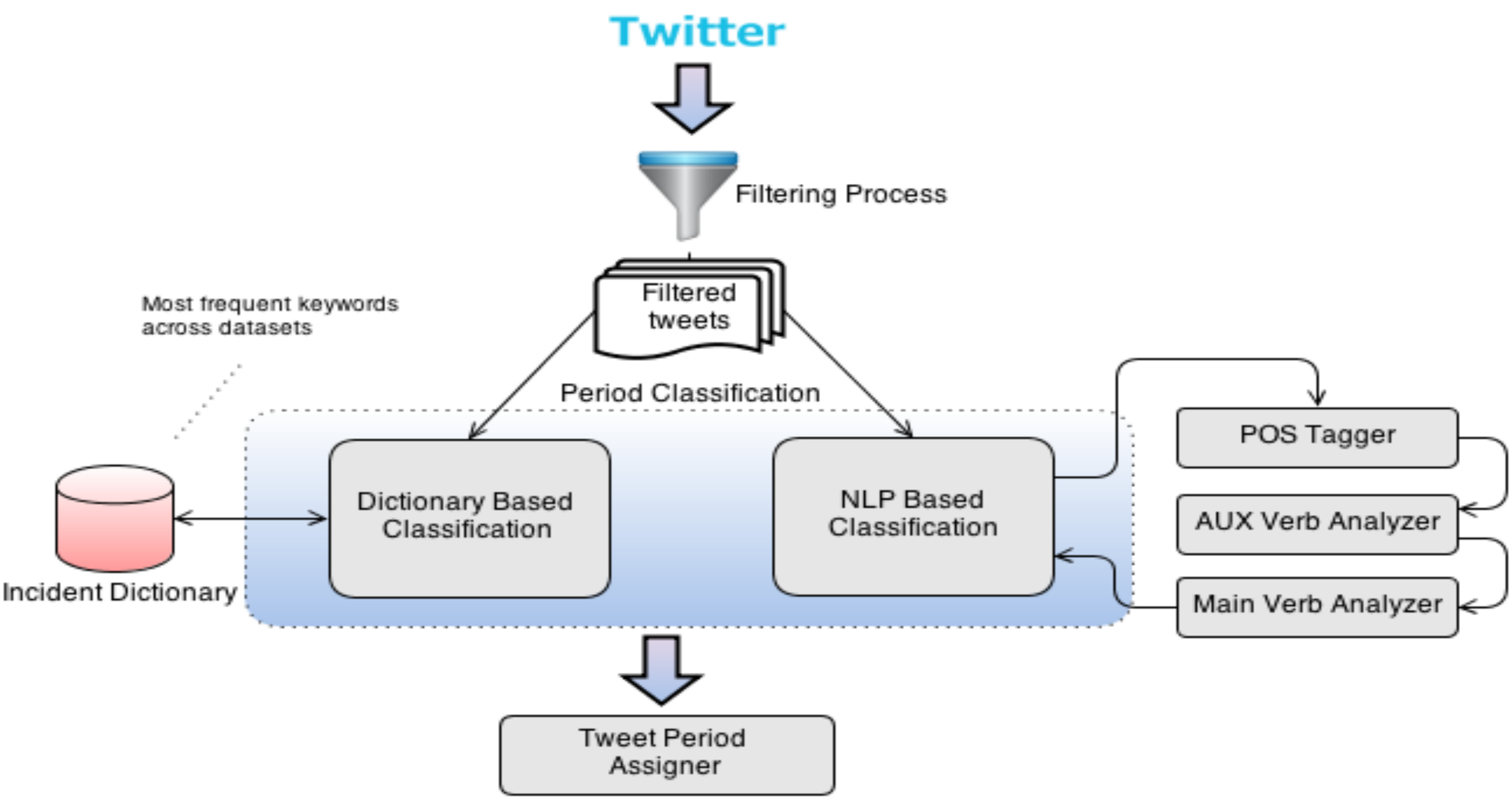
Datasets

Datasets collected using the Twitter Streaming API with appropriate hashtags: those announced by the crisis management authorities at the time of an incident.

1. **Joplin Tornado on May 22, 2011:** 1500 tweets
2. **Haiti Earthquake on Jan 12, 2010:** 1500 tweets
3. **Nesat Typhoon in Philipines on Sep 27, 2011:** 500 tweets

Tweet4Act

Collection -> Filtering -> Period Classification



1. Filtering Process

- **Cleaning: up to 13% are not incident-related**
- **Outlier detection**
  - Normalize message text: remove the “RT @username ” and “@username ” prefixes.
  - Remove duplicate messages after normalization.
  - Remove all terms that appear in less than a fraction s = 0.05 of messages.
  - Run the k-medoid clustering algorithm on each dataset.
  - Discard clusters having a negative number or zero as silhouette coefficient.
  - Select from each cluster the fraction m of messages closer to the medoid.
- **Result = top-m fraction of the most representative messages from each cluster**
- **1,198, 1,167 and 373 unique messages in Joplin, Haiti and Nesat datasets respectively**

Validation on CrowdFlower

- Precision
  - Tweets identified as crisis-related by our method: 498 tweets from Joplin, 250 from Haiti, 200 from Nesat.
  - Each task, which also consists of a set of correctly labeled tweets (i.e., golden data), asked workers to choose if a tweet is crisis-related or not.
  - 3 votes/tweet.
- Recall
  - Random samples of 231, 220, and 244 tweets from Joplin, Haiti and Nesat (before applying the filter).
  - Manually labeled those messages as crisis-related or not.

Filtering Process Validation

Using CrowdFlower

	Joplin (algo)	Joplin (cf)	Haiti (algo)	Haiti (cf)	Nesat (algo)	Nesat (cf)
Unknown	20	10	10	0	3	2
No	9	10	9	0	10	4
Yes	202	478	201	250	231	194

2. Dictionary-based Period Classification

- Compares the words in each message against a dictionary of words known to be present in specific periods of a crisis-incident
- **Most frequent words across datasets**
  - “warning” & “alert” in PRE
  - “now”, “sweeps” in DURING
  - “aftermath”, “donate” in POST

3. NLP-Based Period Classification

Tense of verbs help identify period

(A. Iyengar et al., 2011)

1. If the word is listed in the dictionary, add +1 to the period it is listed under and stop processing that word (i.e., if a verb is in the dictionary, we ignore it below).

2. If the word is an auxiliary verb, add +1 to the period it is associated (e.g., could-PRE, are-DURING, did-POST).

3. If the word is a verb in future/present/past tense, add +0.5 to pre/during/post period, respectively.

4. Sum up scores of each period across all words in the phrase and pick the period with the largest sum.

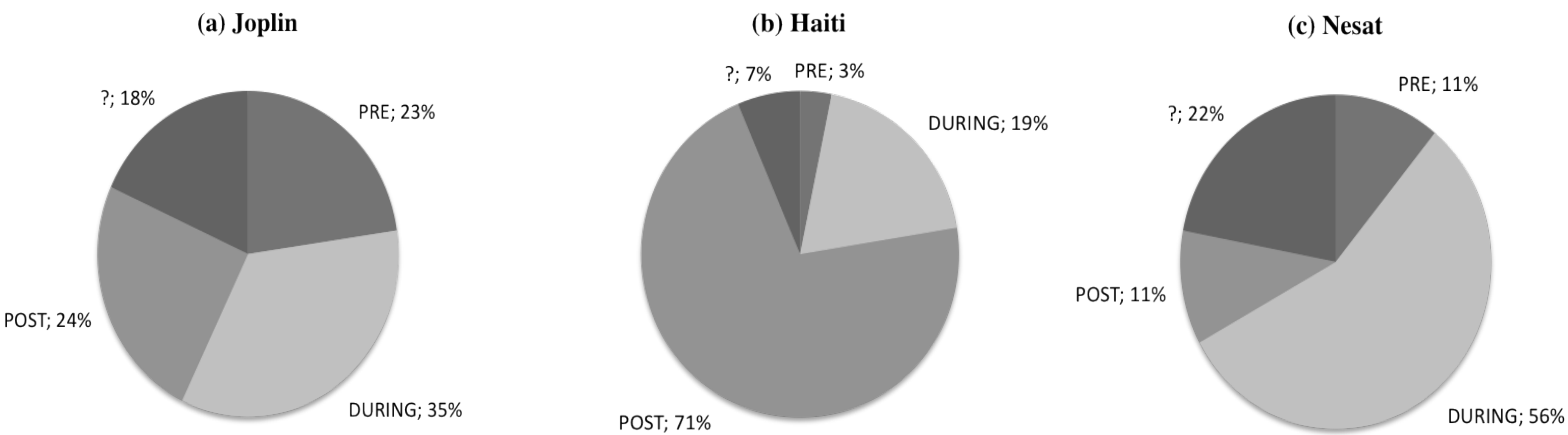
Simple Scoring Example

NFL teams gathering supplies aid for tornado victims in Kansas Missouri (Morning Call) …

- In this message, both words “aid ” and “victim ” are matched in the dictionary for the POST period.
- The verb “gathering ” is in continuous form and contributes to the DURING period.
- In total, the message has +2 score for POST and +0.5 for DURING; hence, it is classified as POST .

Manual Period Classification (labeling)

CrowdFlower period labeling



Performance of Tweet4Act

Period	Tweet4act			SVM			MaxEnt			Tree			RF		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PRE	0.														0
DURING															.48
POST															18
AVG															33
PRE															5
DURING															57
POST															94
AVG															67
PRE															67
DURING															88
POST															.33
AVG	0.														0.63

- *Tweet4act provides better recall over baseline algorithms for all datasets*
- *But, it misses some precision*
  - *Haiti (0.60 against 0.96 by RF)*
  - *Nesat (0.61 against 0.94 by MaxEnt)*

Talk outline

1. **Quick overview of existing crowdsourcing**
2. **An integrated architecture for automatic task assignment**
3. **Online tweet monitoring**
   – Classified tweets into PRE/DURING/POST
4. **Summary and future work**

Summary and Future Work

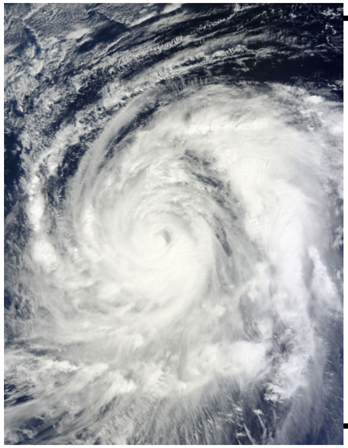
1. Crowdsourcing is a powerful paradigm to help in crisis reporting during and after it happens
2. Implicit reporting serves task awareness
3. Explicit reporting with recurring crowds opens new research opportunities for effective task assignment: to report on crises, to participate in task evaluation
4. Task assignment is effective when skill learning and task evaluation are possible
5. All that is only possible with a general-purpose crowdsourcing platform

Algorithms to Help in Future Natural Disasters

Crowd4U
A Crowdsourcing and Microvolunteering Platform by Academia, for the Universe

Task Results

Research Center for Natural Hazards & Disaster Recovery Niigata University



Real Data on Natural Disasters (Currently, data on Typhoon Wipha in 2013)

Workers

Better Task Design & Algorithms

References

[Boim et. Al] R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W. C. Tan. Asking the right questions in crowd data sourcing. In ICDE, pages 1261–1264, 2012.

[Krager et. Al] D. R. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. CoRR, abs/ 1110.3564, 2011.

[Ramesh et. al.] A. Ramesh, A. Parameswaran, H. Garcia-Molina, and N. Polyzotis. Identifying reliable workers swiftly. Technical report, Stanford University. 2012.

[Ho et. al.] C.-J. Ho and J. W. Vaughan. Online task assignment in crowdsourcing markets. In AAAI, 2012.

Senjuti Basu Roy, Ioanna Lykourentzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, Gautam Das: Crowds, not Drones: Modeling Human Factors in Interactive Crowdsourcing. DBCrowd 2013: 39-42

Senjuti Basu Roy, Ioanna Lykourentzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, Gautam Das : Optimization in Knowledge-Intensive Crowdsourcing. CoRR abs/1401.1302 (2014)

Senjuti Basu Roy, Saravanan Thirumuruganathan, Sihem Amer-Yahia, Gautam Das, Cong Yu: Exploiting group recommendation functions for flexible preferences. ICDE 2014: 412-423

Atsuyuki Morishima, Sihem Amer-Yahia, Senjuti Basu Roy. “Crowd4U: An Initiative for Constructing an Open Academic Crowdsourcing Network”, HCOMP 2014.

Senjuti Basu Roy, Sihem Amer-Yahia,”It Takes a Village” - A Raging Vision to Enable Collaboration in Human Intensive Applications (under review)

Habibur Rahman, Senjuti Basu Roy, Ioanna Lykourentzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, Gautam Das, Optimization in Collaborative Crowdsourcing (under review)

A. Iyengar, T. Finin, and A. Joshi (2011) Content-based prediction of temporal boundaries for events in Twitter. In Proceedings of the Third IEEE International Conference on Social Computing.

K. Starbird, L. Palen, A. Hughes, and S. Vieweg (2010) Chatter on the red: what hazards threat reveals about the social life of microblogged information. In Proceedings of the 2010 ACM conference on Computer supported cooperative work, pages 241–250. ACM.

Latonero, Mark, and Irina Shklovski. "“Respectfully Yours in Safety and Service”: Emergency Management & Social Media Evangelism.” Proceedings of the 7th International ISCRAM Conference– Seattle. Vol. 1. 2010.