

The Role of Information Theory and Queuing Theory in Human Computation Systems

Avhishek Chatterjee

with Lav R. Varshney

University of Illinois at Urbana-Champaign

Research contributions from Michael Borokhovich, Mark Hasegawa-Johnson, Preethi Jyothi, Ravi Kiran Raman, Daewon Seo, Pramod K. Varshney, Aditya Vempaty, and Sriram Vishwanath

Reliability-cost tradeoff in human computation

Natural question in information theory



Is it a picture of a celebrity?

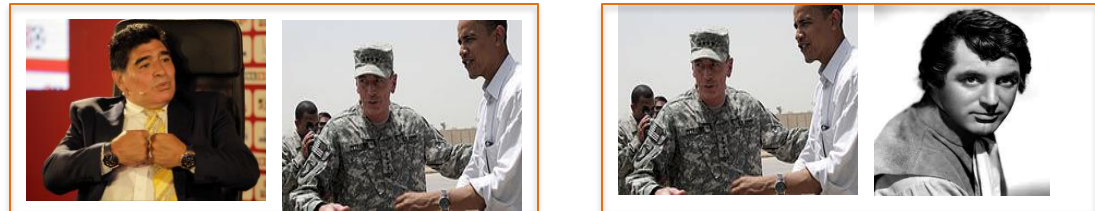
Crowdsourcing: one task to a human each

Significant **error** rate: mistakes, spam, etc.

Add **redundancy**: give one job to multiple humans
(Karger, Oh, and Shah 2011, and more)

number of workers increases, so does cost

Algebraic coding
across tasks



Are both pictures of similar kind (famous or not)?

Better (Vempaty, Varshney, and Varshney 2014), but what is the best?

Parallel to information theory and more

Min. cost for perfect reliability

Capacity of noisy channel

Min. cost for a given reliability

Joint source-channel coding¹

Corrupted or partially known
human performance statistics

Noisy channel with
imperfect side information

No statistics: universal crowdsourcing
(Raman and Varshney, preprint 2016)

Universal communication

Capacity achieving codes are not suited for human computation

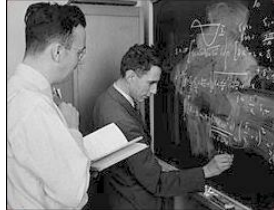
- “constraints” on how to combine tasks and how many
(“even number of celebrity pictures among these 10 pictures?” is not a good question)
- do the fundamental limits change under the
constraints on task combining

(achievability and converse under restriction to a class of Boolean operations: challenging!)

¹Lahouti and Hassibi, NIPS 2016.

Skilled crowdsourcing

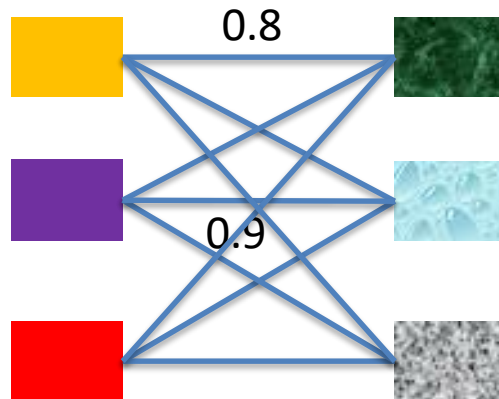
A question at the intersection of information and queuing theory



Not all workers are same for all tasks

- education, profession, nationality, etc.

Arrival process
(stochastic)



Skilled worker
availability
(stochastic)

**Allocate waiting tasks
at regular intervals to
the available workers**

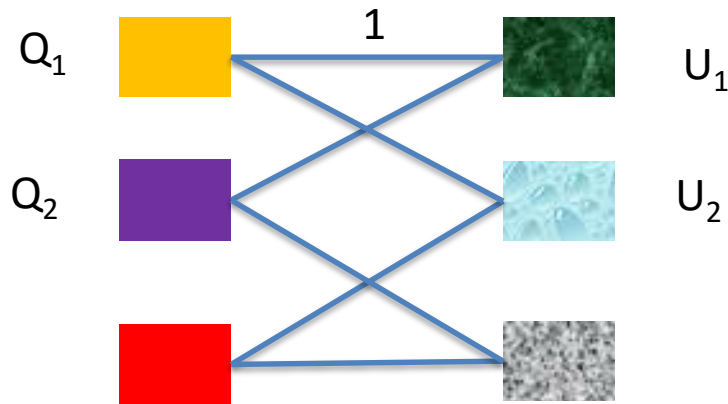
Coding across tasks and task-to-worker matching:
bounded backlog, small delay, low error, and minimal redundancy.

Queue scheduling

Information theory

Allocating arriving tasks to skilled workers

Natural question in queuing systems



- Actions have strong future implications: dynamical system rather than one-shot
- Arrival and availability statistics not known
- Actions must ensure bounded backlog and small delay

An “optimal” policy: comes from queuing dynamics

→ ensures bounded backlog whenever possible

Implementing optimal policy : solve an optimization problem each time

- polynomial computation?
- approximation and its implications on backlog?

Our related work (INFOCOM 2015, 2016): tasks with **multiple steps** and **precedence constraints** - *different from static scheduling with precedence constraint*

Lot to be explored: **low error rate** and **low redundancy**

- intersection of queuing, algorithms, and information theory

Conclusion

- Information theory is natural in handling reliability-cost tradeoff
 - coding schemes with human constraints and fundamental bounds
- Queuing theory is natural in handling task arrival and worker dynamics
 - need to be combined with information theory to capture reliability
- Human computation problems need a completely new kind of union between queuing theory and information theory, e.g.,
 - human performance deteriorates with increasing load
 - best way to send tasks to a worker?*
 - preliminary result in a single worker setting (Chatterjee, Seo, and Varshney, ISITA 2016)
 - multiple worker and task types: **queuing, information theory, and algorithms**



P.S.: Approaches developed based on information and queuing theory are useful in practice

Impact sourcing

- crowdsourcing to empower the underprivileged
- queuing motivated allocation rule works well on data

samaSOURCE

Speech transcription

- information theory enhances use of non-native workers
- mismatched crowdsourcing



Related publications

- [1] M. Borokhovich, A. Chatterjee, J. Rogers, L. R. Varshney, and S. Vishwanath, “Improving Impact Sourcing via Efficient Global Service Delivery,” in *Proceedings of the Data for Good Exchange (D4GX)*, New York, New York, 28 September 2015.
- [2] A. Chatterjee, M. Borokhovich, L. R. Varshney, and S. Vishwanath, “Efficient and Flexible Crowdsourcing of Specialized Tasks with Precedence Constraints,” in *Proceedings of the 2016 IEEE Conference on Computer Communications (INFOCOM)*, San Francisco, California, 10-15 April 2016.
- [3] A. Chatterjee, D. Seo, and L. R. Varshney, “Capacity of Systems with Queue-Length Dependent Service Quality,” in *Proceedings of the International Symposium on Information Theory and Its Applications (ISITA)*, Monterey, California, 30 October – 2 November 2016.
- [4] A. Chatterjee, L. R. Varshney, and S. Vishwanath, “Work Capacity of Freelance Markets: Fundamental Limits and Decentralized Schemes,” in *Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM)*, Hong Kong, 26 April – 1 May, 2015.
- [5] W. Chen, M. Hasegawa-Johnson, N. F. Chen, P. Jyothi, and L. R. Varshney, “Mismatched Crowdsourcing with Clustering-Based Phonetic Projection for Low-Resourced ASR,” to appear in *Proceedings of the 26th International Conference on Computational Linguistics Workshops (COLING 2016)*, Osaka, Japan, 11 December 2016.

Related publications

- [6] M. Hasegawa-Johnson, J. Cole, P. Jyothi, and L. R. Varshney, “Models of Dataset Size, Question Design, and Cross-Language Speech Perception for Speech Crowdsourcing Applications,” *Laboratory Phonology*, vol. 6, no. 3-4, pp. 381-431, October 2015.
- [7] R. K. Raman and L. R. Varshney, “Universal Clustering via Crowdsourcing” arXiv:1610.02276 [cs.HC].
- [8] G. V. Ranade and L. R. Varshney, “To Crowdsourc or not to Crowdsourc?,” in *Proceedings of the 4th Human Computation Workshop (HCOMP)*, Toronto, Canada, 23 July 2012.
- [9] L. R. Varshney, P. Jyothi, and M. Hasegawa-Johnson, “Language Coverage for Mismatched Crowdsourcing,” in *Proceedings of the 2016 Information Theory and its Applications Workshop (ITA)*, San Diego, California, 31 January – 5 February 2016.
- [10] A. Vempaty, L. R. Varshney, and P. K. Varshney, “Reliable Crowdsourcing for Multi-Class Labeling using Coding Theory,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 667-679, August 2014.
- [11] A. Vempaty, L. R. Varshney, and P. K. Varshney, “Reliable Classification by Unreliable Crowds,” in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 26-31 May 2013.