In Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST 2015)
Daegu, Kyungpook, Republic of Korea

Investigating the "Wisdom of Crowds" at Scale

Alok Shankar Mysore⋆
PES Institute of Technology
alok.shankar@pesit.pes.edu

Vikas S Yaligar‡
National Institute of Technology Karnataka
vikasyaligar@ieee.org

Imanol Arrieta Ibarra
Stanford University
imanol@stanford.edu

Camelia Simoiu
Stanford University
csimoiu@stanford.edu

Sharad Goel
Stanford University
sgoel@stanford.edu

Additional Authors[*]
Various Institutions[†]

ABSTRACT
In a variety of problem domains, it has been observed that the aggregate opinions of groups are often more accurate than those of the constituent individuals, a phenomenon that has been termed the "wisdom of the crowd." Yet, perhaps surprisingly, there is still little consensus on how generally the phenomenon holds, how best to aggregate crowd judgements, and how social influence affects estimates. We investigate these questions by taking a meta wisdom of crowds approach. With a distributed team of over 100 student researchers across 17 institutions in the United States and India, we develop a large-scale online experiment to systematically study the wisdom of crowds effect for 1,000 different tasks in 50 subject domains. These tasks involve various types of knowledge (e.g., explicit knowledge, tacit knowledge, and prediction), question formats (e.g., multiple choice and point estimation), and inputs (e.g., text, audio, and video). To examine the effect of social influence, participants are randomly assigned to one of three different experiment conditions in which they see varying degrees of information on the responses of others. In this ongoing project, we are now preparing to recruit participants via Amazon's Mechanical Turk.

Author Keywords
Crowdsourcing; online experiment; crowd consensus.

ACM Classification Keywords
H.5.m. Economics: Experimentation Design

INTRODUCTION
At a 1906 county fair, the statistician Francis Galton watched as eight hundred people competed to guess the weight of an ox. He famously observed that the median of the guesses, 1,207 pounds, was, remarkably, within 1% of the true weight [1].

Simple aggregation—as in the case of Galton's ox competition, or voting in democratic elections—has been shown to be a surprisingly powerful technique for prediction, inference, and decision-making. Over the last century, there have been dozens of studies that examine this wisdom of crowds effect. For example, crowd judgements have been used to identify phishing websites [5], answer general knowledge questions [4], and forecast weather-related events [2]. In these applications, a wide variety of aggregation methods have been considered, ranging from standard measures, such as the mean and median, to more specialized, domain-specific techniques, such as those based on cognitive models of decision making [3]. However, given the diversity of experimental designs, subject pools, and analytic methods employed, it has proven difficult to compare studies and extract general principles. It is thus unclear whether these documented examples are a representative collection of a much larger space of tasks that exhibit a wisdom of crowds phenomenon, or conversely, whether they are highly specific instances of an interesting, though ultimately limited occurrence.

[*]Ramesh Arvind⋆, Chiraag Sumanth⋆, Arvind Srikantan⋆, Bhargav HS⋆, Mayank Pahadia‡, Tushar Dobhal‡, Atif Ahmed‡, Mani Shankar‡, Himani Agarwal, Rajat Agarwal, Sai Anirudh-Kondaveeti, Shashank Arun-Gokhale, Aayush Attri, Arpita Chandra, Yogitha Chilukuri, Sharath Dharmaji, Deepak Garg, Naman Gupta, Paras Gupta, Glincy Mary Jacob, Siddharth Jain, Shashank Joshi, Tarun Khajuria, Sameeksha Khillan, Sandeep Konam, Praveen Kumar-Kolla, Sahil Loomba, Rachit Madan, Akshansh, Vidit Mathur, Bharat Munshi, Mohammed Nawazish, Venkata Neehar-Kurukunda, Venkat Nirmal-Gavarraju, Sonali Parashar, Harsh Parikh, Avinash Paritala, Amit Patil, Rahul Phatak, Mandar Pradhan, Abhilasha Ravichander, Krishna Sangeeth, Sreecharan Sankaranarayanan, Vibhor Sehgal, Ashrith Sheshan, Suprajha Shibiraj, Aditya Singh, Anjali Singh, Prashant Sinha, Pushkin Soni, Bipin Thomas, Lokesh Tuteja, Kasyap Varma-Dattada, Sukanya Venkataraman, Pulkit Verma, Ishan Yelurwar

[†]Jaypee Institute Of Information Technology; BITS; National Institute of Technology Karnataka, Indian Institute of Technology Delhi; PES Institute of Technology; International Institute of Information Technology; BMS College of Engineering; Bhagwan Parshuram Institute of Technology; Indian Institute of Technology Guwahati; College of Engineering; College of Engineering Chengannur; Bharati Vidyapeeth's College Of Engineering; Maharaja Agrasen Institute of Technology; Cluster Innovation Centre; Rajiv Gandhi University of Knowledge Technologies; Indraprastha Institute of Information Technology

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
UIST'15, November 8-11, 2015, Charlotte, NC., USA.
Copyright © 2015 ACM.
DOI: http://dx.doi.org/10.1145/2815585.2815725

EXPERIMENT DESIGN

To systematically investigate the wisdom of crowds, we developed a modular online experiment that presents participants with 1,000 questions drawn from 50 domains. Each domain includes 20 questions on a specific topic. For example, one domain tests individuals' knowledge of geography, and asks them to identify 20 countries from their silhouettes.

Domains include tests of explicit knowledge, tacit knowledge, and prediction ability. Explicit knowledge refers to that which can be readily articulated, accessed and communicated to others. In this category, we test general knowledge, spatial reasoning, and popular culture. Tacit knowledge refers to the type of knowledge that is difficult to transfer to another person and is often gained through experience. In this category, we test emotional intelligence, knowledge of cultural norms, and foreign language skills. Prediction domains include election outcomes, and box office success of upcoming movies. The domains span four different types of media (text, image, video and audio) and four question types (point estimate, binary, multiple choice, and ranking).

When participants begin a new domain, they are asked to estimate their ability relative to the crowd, "Out of 100 people answering these questions, where do you think you will place? We also ask participants to report their level of confidence (on a 5-point scale) on each question (see Figure 1). With this information, we seek to more efficiently aggregate answers by accounting for self-assessed expertise.

If participants receive information about the guesses of others, that might plausibly yield far better answers than those obtained via independent responses. But there is also a worry that such social influence would result in herding behavior, which in turn could decrease collective performance [4]. To investigate the effects of social influence, participants are randomly assigned to one of three experimental conditions. In the first group, participants are shown the previous five answers to the question; in the second group, they see the previous five answers with the highest self-reported confidence. The third group is the control, in which participants do not receive any information about others' responses.

To encourage participation and engagement, we incorporate two competitive elements into the design of the platform. A timer is included which is set to expire in 30-45 seconds (depending on the question type). Second, we adopt a theme for the platform that lends a social and collaborative feel to the experiment. Namely, participants progress through a series of bee-like-avatars as they complete domains, ultimately becoming the "queen bee" if they complete all the questions.

COLLABORATIVE DESIGN & IMPLEMENTATION

Given the scale of the experiment, its design and implementation was conducted collaboratively by a group of approximately 100 undergraduate students in India led by two graduate students at Stanford, as part of Stanford's Aspiring Researchers Challenge (`http://aspiringresearchers.soe.ucsc.edu`). 54 of these students are listed as co-authors on this submission based on their contribution to the project. This unconventional approach—in which we leveraged the wisdom of crowds to study the wisdom of crowds—allowed us to analyse a much larger and more diverse set of tasks than would be possible with traditional methods.

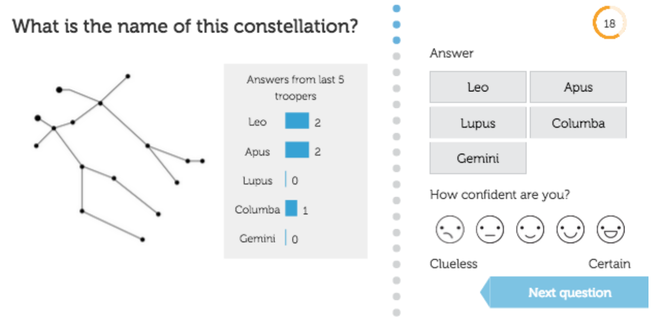

Figure 1. Sample multiple choice question with social condition.

We started by conducting a collaborative literature review, finding and summarizing over 100 related papers on the subject. We then collectively brainstormed nearly 200 ideas for question categories, of which 50 were ultimately selected to be included in the experiment. Creating a domain requires finding a suitable corpus of questions, and, for certain domains, creating and editing the necessary movies, audio, and images. We accomplished these tasks by splitting into 23 smaller teams of 2-5 students. Finally, the experiment platform was fully implemented by two specialized teams of students who focused on front-end * and back-end design ‡.

PROJECT STATUS & FUTURE WORK

A preliminary version of the experiment is available at `http://wisdomofcrowds.stanford.edu`. Having completed this first phase of the project, we are now working to integrate the platform with Amazon's Mechanical Turk in order to recruit participants. After running a series of pilot studies, we expect to launch a large-scale version of the experiment, with a goal of eliciting 100 responses for each of 1,000 questions in three different social information conditions, for a total of 300,000 responses.

REFERENCES

1. Galton, F. Vox populi (the wisdom of crowds). *Nature 75* (1907), 450–451.
2. Hueffer, K., Fonseca, M. A., Leiserowitz, A., and Taylor, K. M. The wisdom of crowds: Predicting a weather and climate-related event. *Judgment and Decision Making 8*, 2 (2013), 91–105.
3. Lee, M. D., Zhang, S., and Shi, J. The wisdom of the crowd playing the price is right. *Memory & cognition 39*, 5 (2011), 914–923.
4. Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences 108*, 22 (2011), 9020–9025.
5. Moore, T., and Clayton, R. Evaluating the wisdom of crowds in assessing phishing websites. In *Financial Cryptography and Data Security*. Springer, 2008, 16–30.