Crowd Research: Open and Scalable University Laboratories

Rajan Vaish[1], **Snehalkumar (Neil) S. Gaikwad**[2], **Geza Kovacs**[1], **Andreas Veit**[3], **Ranjay Krishna**[1], **Imanol Arrieta Ibarra**[1], **Camelia Simoiu**[1], **Michael Wilber**[3], **Serge Belongie**[3], **Sharad Goel**[1], **James Davis**[4], **Michael S. Bernstein**[1]
[1]Stanford University, [2]MIT Media Lab, [3]Cornell Tech, [4]UC Santa Cruz
{rvaish, msb}@cs.stanford.edu

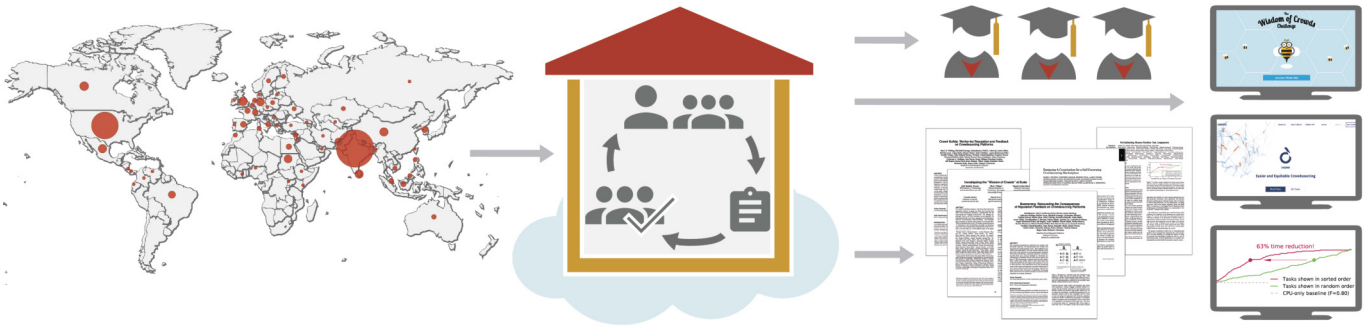

Figure 1. We present a crowdsourcing technique that has enabled access to research experiences for over 1,500 people from 62 countries. Participants achieve upward educational mobility while creating research systems and co-authoring papers at top-tier ACM venues such as CSCW and UIST.

ABSTRACT
Research experiences today are limited to a privileged few at select universities. Providing open access to research experiences would enable global upward mobility and increased diversity in the scientific workforce. How can we coordinate a crowd of diverse volunteers on open-ended research? How could a PI have enough visibility into each person's contributions to recommend them for further study? We present Crowd Research, a crowdsourcing technique that coordinates open-ended research through an iterative cycle of open contribution, synchronous collaboration, and peer assessment. To aid upward mobility and recognize contributions in publications, we introduce a decentralized credit system: participants allocate credits to each other, which a graph centrality algorithm translates into a collectively-created author order. Over 1,500 people from 62 countries have participated, 74% from institutions with low access to research. Over two years and three projects, this crowd has produced articles at top-tier Computer Science venues, and participants have gone on to leading graduate programs.

UIST 2017, October 22–25, 2017, Quebec City, QC, Canada

ISBN 978-1-4503-4981-9/17/10…$15.00
DOI: https://doi.org/10.1145/3126594.3126648

Author Keywords
Crowdsourcing; citizen science; crowd research

ACM Classification Keywords
H.5.3. HCI: Group and organization interfaces.

INTRODUCTION
Scientific research remains the domain of the privileged few. Those blessed with the socioeconomic opportunity to attend prestigious universities can gain access to research experiences that support open-ended inquiry, train scientific minds and launch careers [64]. Unfortunately, these opportunities remain out of reach for the vast majority of people worldwide [5, 56, 3]. Such people may have the creativity, insight, and work ethic to produce major achievements, but lack access to the opportunity. The result is an ecosystem that systematically underrepresents minorities and developing regions, and a literature that overlooks their diverse perspectives.

Providing open access to research experiences would open new channels for upward educational and career mobility worldwide. However, how can a principal investigator such as a faculty member or research scientist coordinate an entire crowd of diverse people? If the goal is to give participants full breadth to demonstrate creativity, solve unanticipated challenges, and guide the project's direction — not to reduce them to mechanical research assistants — no general techniques yet exist. Citizen science efforts have pursued protein folding [10, 47], scientific dataset labeling [71, 42], math proofs [13], and experiment replication [9], but these projects required

pre-defined, static goals rather than allowing participants to iteratively guide the research goal. In addition, sheer scale prevents a principal investigator from having full visibility into each participant's contributions, threatening their ability to recommend participants for further study.

This paper describes *Crowd Research*, a crowdsourcing technique that enables open access for a global crowd to work together on research under a principal investigator (PI). Crowd Research participants collaborate online as one large team to brainstorm research ideas, execute solutions, and publish scholarly articles — a university laboratory at massive scale. To facilitate open access, we introduce a crowdsourcing technique that comprises weekly cycles of contribution, synchronous collaboration, and peer assessment to produce each next week's iterative goal. A suite of systems carries research from ideation to execution, including brainstorming, engineering, design, analysis and paper writing. For the PI, Crowd Research offers a chance to convene hundreds or thousands of people on a single massive project, enabling research achievements at a scale that is rare today.

To enable upward career and educational mobility, Crowd Research must provide contributors with credible evidence of their impact. However, a PI cannot easily disaggregate participants' interdependent contributions, and may not have centralized visibility into each participant's work. We thus introduce a decentralized credit system where participants allocate credit to each other. This allocation process creates a weighted directed graph, enabling our system to apply a graph centrality algorithm to determine a collectively-created author order for publication and the PI's recommendation letters.

Crowd Research has so far brought together over 1,500 participants from six continents, 74% of whom come from universities ranked below 500 in global research activity and influence by Times Higher Education [1]. It has included three different research projects with four PIs from Stanford, UC Santa Cruz and Cornell — ranging from human-computer interaction (HCI) to data science to artificial intelligence (AI). These projects produced crowd-authored papers that have been accepted to top-tier Computer Science venues including ACM UIST [22] and ACM CSCW [80]. Participants have leveraged their contributions to receive recommendation letters from PIs. Despite having a median of zero other letter writers from institutions ranked above 500 worldwide, participants have been admitted for further study at undergraduate and graduate programs at universities such as Stanford, UC Berkeley, Carnegie Mellon, and MIT.

Our contributions span a crowdsourcing technique for coordination of open-ended, long-term and complex goals; a decentralized method for allocating credit; and an analysis of a two year, large-scale deployment of the method. To follow, we position Crowd Research in related work, describe the technique, and analyze our deployment and limitations.

RELATED WORK

Research experiences are not just authentic practice: they impact upward mobility. Engaging in research increases the probability of enrolling in STEM graduate programs [19], both for professional degrees [49] and PhDs [84]. Research experiences also increase interest in STEM careers [64] and increase a student's likelihood of using faculty recommendations for jobs [27]. Other improvements accrue to technical skills, interpersonal skills, and scientific literacy [43, 14].

The size of Crowd Research and the diversity of its membership offers new opportunities for science and engineering research. Having many people brings a diversity of ideas [41, 50, 82]. Diversity arguably brings even greater benefits: diverse problem solvers outperform groups of high-ability problem solvers [30], and diverse perspectives unearth hidden assumptions and yield more active and effortful thought [26]. Our work synthesizes these benefits by drawing on a diverse worldwide population, and applying them toward open-ended research goals. In doing so, it trades off the expertise that most research projects can assume of their participants (e.g., graduate-level coursework), and instead uses peer assessment [39], the web, and lectures to provide on-demand training.

Research experiences have traditionally been one-on-one cognitive apprenticeships [31]. Providing mentorship is a nontrivial time commitment for faculty, which often limits who can be mentored [83]. More critically, universities which produce the world's most-cited research are concentrated in North America and Europe [1], far from the world's largest and developing population centers. Crowd Research introduces techniques to bring the benefits of research experiences to a far larger group.

Online Access to Training and Science

Crowd Research draws lessons from online education and citizen science, each of which expands access to opportunities that are typically only available within universities. MOOCs democratize access to online learning opportunities [15], offering an attractive template for Crowd Research. Unfortunately, MOOCs are especially likely to leave behind people from less developed areas [37], and taking a MOOC does not translate to upward career mobility [16]. Crowd Research builds on these efforts by directly encouraging authentic practice and enabling calibrated assessment for upward mobility.

Citizen science enables members of the public to contribute to research [68]. Typically these projects predefine the goal of the project and the method of contribution, and participants contribute by filling out the "rows" of the desired dataset. For example, projects engage volunteers to upload bird locations on eBird [71], take tests on LabInTheWild [61], and label galaxies on Zooniverse [11]. Other projects give participants more freedom in how they answer a research question, for example, crowdsourced math proofs in the Polymath project [13] and protein folding in Foldit [33]. Crowd Research represents a rarer third category, *co-created projects*, where participants are involved not just in data collection and execution but also in the conception and ongoing evolution of the research [4, 57, 54]. Crowd Research is unusual even in this class of projects because participants own the whole research arc, rather than one focused part.

While citizen science has succeeded in engaging with participants worldwide, it has struggled to close the access gap. For example: (i) Zooniverse participants tend to be from highly

educated countries [59]; (ii) Nearly all Polymath participants were faculty or Ph.D. students, 86% had published papers, and only one was known to be female [13]; and (iii) OpenStreetMap contributors are 96% male, with three-quarters holding a postgraduate degree [6]. However, some citizen science projects have explicitly attempted to incorporate marginalized communities [70]. Crowd Research reaches a global audience via diversified recruiting and provides direct incentives for upward mobility such as paper authorship and recommendation letters. However, Crowd Research cannot yet overcome internet infrastructure and language limitations.

Coordinating Research and Crowds

To enable crowds to engage in collaborative research, Crowd Research extends work from CSCW and social computing. Authentic tasks (e.g., [73]) and feedback (e.g., [18, 40]) are both critical elements to improvement. Within traditional laboratory environments, pair working sessions can support knowledge transfer [52], and agile research studios can scale mentorship per PI to about twenty students [83]. Crowd Research operates at a much larger scale and with more diverse participants. This requires different approaches, in particular fewer team-based agile methods and more structured, pre-defined milestones. Decisionmaking must also become more decentralized, e.g., via peer assessment, decentralized credit allocation, and DRIs.

Crowdsourcing techniques increasingly aim to support complex outcomes [36, 35]. These systems can now support goals ranging from software engineering [45, 7] to writing [53, 74]. Crowd Research shares some characteristics with this work, organizing the crowd into expertise-based teams [62] and hierarchical structures [78] that can adapt as the crowd proceeds. Unlike prior work, Crowd Research is designed to train participants, so it introduces explicit peer feedback and direct engagement with the PI as a leader.

Determining Credit

To aid upward mobility, Crowd Research must provide assessments of participants' contributions that they can leverage for school and job applications. One strategy for determining author order is to alphabetize. However, women receive less credit than men in alphabetical author orders [65]. A second strategy is to publish as a single joint author, as in "DHJ Polymath" in the Polymath Project [25]. However, a joint name does not provide strong signals for participants to leverage for recommendations. Firms such as Quirky and Assembly offer credit for pre-defined contribution categories, e.g., 1% for coming up with the product's name. However, research is an iterative process where it is not always clear which contributions will wind up being influential. So, we develop a new, decentralized credit approach.

Prior work has studied algorithmic ranking schemes, for example hubs and authorities [38] and PageRank [55]. Similar schemes have been applied to curation [81] and citations in order to determine influence [72, 17, 67]. However, these approaches all assume the existence of a network, which Crowd Research does not have. So, Crowd Research introduces a technique that allows all participants to have a say in the eventual allocation of credit, translating credit into a graph problem.

Figure 2. Crowd Research comprises weekly meetings to discuss the project, milestones to submit concrete progress, and peer assessment to identify top submissions.

CROWD RESEARCH

Crowd Research (Figure 2) introduces a crowdsourcing technique to enable worldwide access to research experiences without overwhelming a PI. In this section we present the approach in detail, oriented around (i) how Crowd Research coordinates large groups of participants, (ii) what systems enable collaboration and scholarly outcomes, and (iii) how it enables upward mobility through a decentralized credit system.

Coordination strategy and process

Crowd Research enables thousands of people online to coordinate joint progress on an open-ended research effort. Prior work has often pre-structured the crowd's contributions — for example providing interfaces for folding proteins [33] — because the goal and the tools needed for success could be defined a priori. Many researchers have eschewed crowdsourcing for exactly this reason: "the process of discovery can be highly uncertain, iterative, and often serendipitous", making the reduction to a crowdsourcing process "hard to imagine" [46]. So, Crowd Research introduces an iterative crowdsourcing technique based on milestones and peer assessment that allows the effort to iterate and adapt over time.

We will refer to the roles of *PI (principal investigator)*, who advises the project; *RA (research assistant)*, who supports logistics, and *participants*, who are members of the crowd. The PIs' motivation was to tap into a diversity of perspectives, mentor far more students than they normally could in their careers, and try out more ambitious projects than typical in their labs. Each PI recruited two RAs to help. The RAs put in a few hours per week — in no cases were these projects the RAs' primary research — mainly helping with onboarding, analyzing top submissions, and answering logistical questions.

Open call recruitment

The first step in Crowd Research is to recruit a crowd. Crowd Research opens with a global online call inviting people to join one of the available posted projects. A public web page describes the opportunity, the PIs involved, and their institutions. We shared this page via social media on Twitter and Facebook groups, cold emails to faculty at international universities, and publicly accessible mailing lists. Interested participants have several weeks to sign up alone or in teams. While selective recruitment is possible, to maximize accessibility we accept all participants who signed up, and create accounts for them on our collaboration platforms.

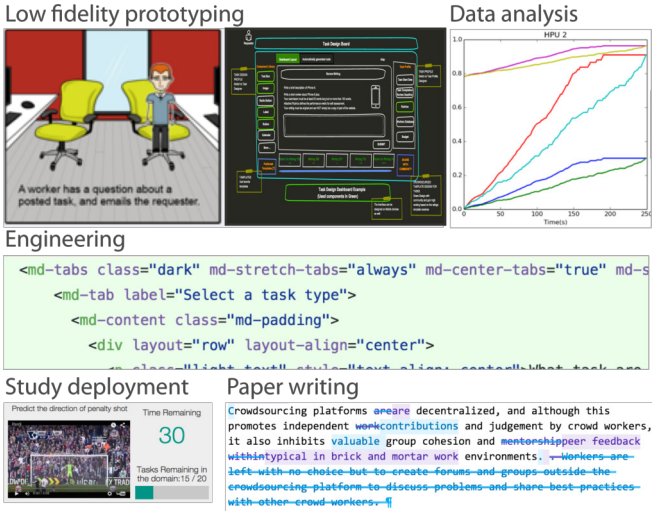

Figure 3. Milestones included prototyping, engineering, and writing.

We launched three different Crowd Research projects, helping us understand how Crowd Research differs across different PIs and research areas. The PIs chose and seeded initial ideas for the projects, much like an initial idea might be seeded with a traditional graduate student. Each PI later developed the idea in collaboration with the crowd.

First, the *human-computer interaction (HCI)* project, led by Prof. Michael Bernstein at Stanford University, set out to create a new paid crowdsourcing marketplace a la Amazon Mechanical Turk. In current crowdsourcing marketplaces, workers feel disrespected, and requesters do not trust the results they receive [32, 51]. The HCI project works on designing, engineering, and studying a new crowdsourcing marketplace, Daemo, to improve work quality and give workers governance of the platform. Second, the *computer vision* project, led by Prof. James Davis at UC Santa Cruz and Prof. Serge Belongie at Cornell Tech, seeks to improve visual classification accuracy. Integrating off-the-shelf machine classifiers with paid crowds is challenging [63]. This project explores strategies to increase accuracy and decrease cost under this setting. Third, the *data science* project, led by Prof. Sharad Goel at Stanford University, seeks to design and run the world's largest "wisdom of crowds" experiment. There is still little consensus on how generally the wisdom of crowds phenomenon holds, how best to aggregate judgments, and how social influence affects estimates. This project tests these boundary conditions by collectively designing and developing 1,000 different prediction tasks in 50 subject domains, and launching them as a large-scale meta-experiment.

Milestone submission

Each week, the PI identifies a concrete goal for the crowd, called a *milestone*. Milestones are scoped at five to ten hours of work per week. Past milestones have included (i) participating in a needfinding interview with Mechanical Turk users, (ii) engineering an experimental scaffold for experiments, (iii) proposing experimental designs, (iv) implementing a proposed algorithm from a previous week, and (v) brainstorming iterations of the research idea based on feedback

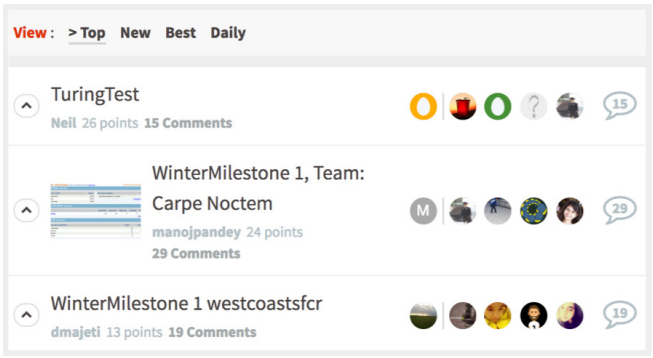

Figure 4. Participants view each others' milestone submissions, leave comments, and upvote promising ideas.

from the PI (Figure 3). For example, one milestone in the HCI project was focused on needfinding, and involved reading papers, joining a panel interview with workers and requesters, and then synthesizing insights. Each project maintains a wiki where the PI or an RA uploads details for all milestones.

Participants work in parallel during the week, individually or in teams, to complete the milestone. Since the process works on a weekly cycle, participants have about six days to complete each milestone. The Slack group chat platform operates as a brainstorming and discussion room for participants where they can interact with each other, help each other, and ask questions. Slack can become quite busy. To manage it, some channels (e.g., #announcements) are low traffic and intended to be read in their entirety. Others are busy and scoped narrowly to a milestone area (e.g., #design, #engineering). Participants often create ad-hoc channels for each milestone and other interest-based channels (e.g., #highschoolers, #machinelearning) to meet other like-minded participants. This helped participants selectively follow relevant conversations without getting overwhelmed.

In the early phases of the project, milestones were limited to one goal each week. However, it soon became clear that the crowd brought many different skills to the projects, and some participants would wait for weeks for their skills to be applicable. So, we began to allow multiple parallel milestones each week, enabling participants to self-organize and select which ones to complete. For example, one week's milestones might include creating interaction mockups for designers, a front-end feature implementation for AngularJS engineers, and a back-end feature implementation Django/Python engineers.

At the end of the week, teams submit their milestones to a peer assessment system. Participants create a page on the wiki containing their milestone submission, and submit that link.

Peer assessment

At this point, there are a large number of submissions to the milestone — far too many for the PI to read and synthesize. They vary greatly in quality, content, and coherence. The next stage of Crowd Research harnesses peer assessment to give feedback on the submissions and provide a rough ranking so that the PI can concentrate on the most promising ideas. The peer assessment process is open for one day, with a cutoff for feedback a few hours before the weekly team meeting.

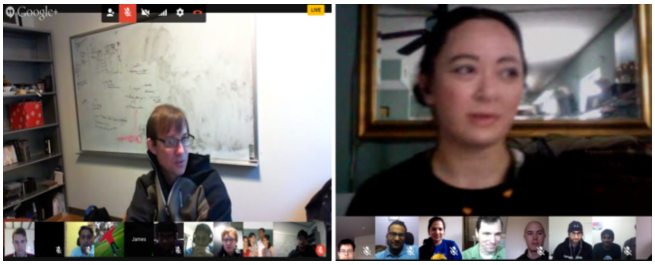

Figure 5. Weekly video meetings on YouTube Live include participants who submitted highly-rated milestones.

Our peer assessment system functions similarly to a social aggregator such as Reddit. Once the submission deadline has passed, the crowd can look at each others' submissions, leave comments and upvote strong submissions. The PI and RAs choose a default sort for the system: e.g., most upvotes at the top to encourage feedback on promising candidates, or fewest comments at the top to encourage diversity. While we initially experimented with a system that randomized a double-blind assignment, anonymous feedback was needlessly negative and evaluative. Instead, we shifted to a system (Figure 4) where participants' names, submissions and feedback were all public, participants chose to give feedback to, prompting a more positive environment.

At the conclusion of the feedback period, the process has called out some of the most interesting and inspirational submissions. This set is small enough for the PI or RAs to collate. They read these submissions and use them as the basis for discussion in the weekly team meeting. While an upvoting process is not perfect [24], it succeeds at separating the insightful submissions from the submissions that are ill-formed, incomplete, or do not display enough understanding.

Weekly video meeting

Once the PI and RAs read the filtered submissions, they discuss next steps with the crowd, much like a PI would with a traditional graduate student. Crowd Research concludes its weekly cycle with a live video meeting broadcast via YouTube Live (Google Hangouts on Air).

This one-hour video meeting (Figure 5) is scheduled so that as many participants as possible can attend: morning in North America enables Europe, China and India to join. The call is streamed and archived automatically on YouTube for anyone who cannot join. The PI or RAs invite participants with highly-rated submissions to join the call live and explain their ideas to the rest of the crowd and the PI. Since the group video meeting has a maximum capacity, other participants join the stream and contribute via a #meetings channel on Slack. The PI informally rotates the invitations to join the call each week to ensure a distribution of nationalities, genders, and backgrounds.

The video call re-aligns all participants, whose ideas may have diverged in many directions during the week. First, the PI begins with a *Rewind* that recaps the last week's goal and progress in case participants missed a week. Second, the PI or RAs share a synthesis of that week's highly-rated milestone submissions. Participants live on the call explain their submissions, and other participants contribute via Slack, which the PI echoes into the live call. The result of this process is that all participants, even those without highlighted submissions, reset their understanding to the "argmax" of the best work.

The PI uses the last few minutes to lay out the next week's milestone, which goes live on the wiki after the call. Participants then begin working, and the process repeats.

Leadership, training, publishing

Complementing the weekly process, we developed training and leadership structures to help focus the crowd's efforts.

DRIs and ad-hoc teams

In the early weeks of the project, divergent ideation is essential for brainstorming research ideas, proposing algorithms and experimental designs, generating design mockups, piloting software or studies, and initial writing. However, some efforts require convergence and collective execution. For example, engineering a feature, making decisions on many different proposed directions, and writing a paper all require that participants work interdependently and collaboratively.

For interdependent milestone goals, the PI empowers a Directly Responsible Individual, or *DRI* [44, 62]. DRIs either self-nominated or were nominated by the PI to lead a milestone based on consistently high-rated milestone submissions. DRIs take charge of a milestone for that week, coordinating any participants who want to contribute to that milestone. They organize ad-hoc video meetings, delegate, and make decisions, summarizing the results in a team submission for the milestone. Being a DRI is a recognition of a participant's contributions, empowering them to have more control over decision-making — and scaling the coordination process. Over time, DRIs overtook many of the RAs' responsibilities, and the process became more community driven.

Training and enrichment

Participants do not all enter the project with sufficient knowledge of the domain. PIs have two main routes for training participants: milestones and video meeting lectures. First, with milestones, a PI can ask participants to read papers and submit commentaries, much like a traditional graduate course, to ensure that participants have the research grounding. Likewise, a milestone might include completing a coding tutorial, or participating in an experiment in order to understand how to design one. Second, with video meeting lectures, the PI can reappropriate an overview lecture from an offline class to teach the crowd a concept that will be important for the research. For example, the PI might give a lecture on one style of computer vision algorithms.

Crowd research also offers an opportunity to connect participants with famous researchers who can serve as inspirational role models. These video meetings so far have included computer scientists such as Andrew Ng (Professor at Stanford and Co-Founder of Coursera), Peter Norvig (Google Research), and Anant Agarwal (MIT and EdX).

Paper writing

Massively collaborative paper writing [76] requires that the crowd integrate its work into academic prose. By this phase of the project, typically a set of DRIs have arisen who can

lead writing of sections of the paper. The PI identifies model papers whose argument structure are similar to the envisioned paper, and then the crowd begins weekly writing iterations on the introduction and framing of the paper. The writing itself happens via collaborative editors. In initial paper-writing efforts, participants were hesitant to overwrite each others' prose. Transitioning to a platform that supported commenting and tracking changes (e.g., Google Docs) was key in making participants feel comfortable contributing. The PI gives feedback as they would on a student's paper. Eventually, the group submits their paper for publication.

Designing for a decentralized credit system

For Crowd Research to deliver on its promise of upward mobility, it must generate credible signals of participants' level of contribution. Participants go on to apply to graduate schools and jobs, and request recommendation letters from the PI. When applicants come from traditionally under-represented areas, the PI's recommendation letter may be the only personal assessment that the company or admission committee trusts. So it is critical for the PI to be able to specify clearly: did a given participant act in a support role, or did they take a leading role in driving the project? However, with interdependent work on open-ended research, the PI may not have visibility into everyone's contributions, and it may be challenging to disaggregate them a priori.

Typical solutions to credit assignment in research and practice are *centralized*: they rely on a single supervisor, or a small number of peers on the team, to make the assessment. For example, the lead researcher often determines author order for all collaborators on a paper, and a worker's supervisor determines the performance review. However, no single person can have a full view of another's contributions [23]. So, not only can centralized credit assignment not scale to Crowd Research, but the PI would be an inaccurate assessor for many participants.

In this paper, we introduce a *decentralized* credit system, which considers every participant's opinion in determining credit [1]. To create a decentralized credit system, we transform the credit problem into a graph problem. This transformation allows us to draw on the tools of network science. In this approach, all participants provide peer assessments about others they have interacted with, and the algorithm aggregates these assessments to determine a final evaluation.

However, with any credit system, it is important to consider possible strategic behavior to influence author order. Not all strategies are malicious: some participants only interacted with a small percentage of the crowd. Graph centrality algorithms can help correct for these strategies. The most common form of manipulation, concentrating all credit within a small subgroup, is similar to a link ring or affiliate networks in web search. Another attack is to strategically direct credit toward others who are likely to send credit back to you — a quid-pro-quo strategy seen in 360-degree reviews [75]. Our strategy must compensate for these behaviors.

[1]The system is available at http://creddit.stanford.edu.

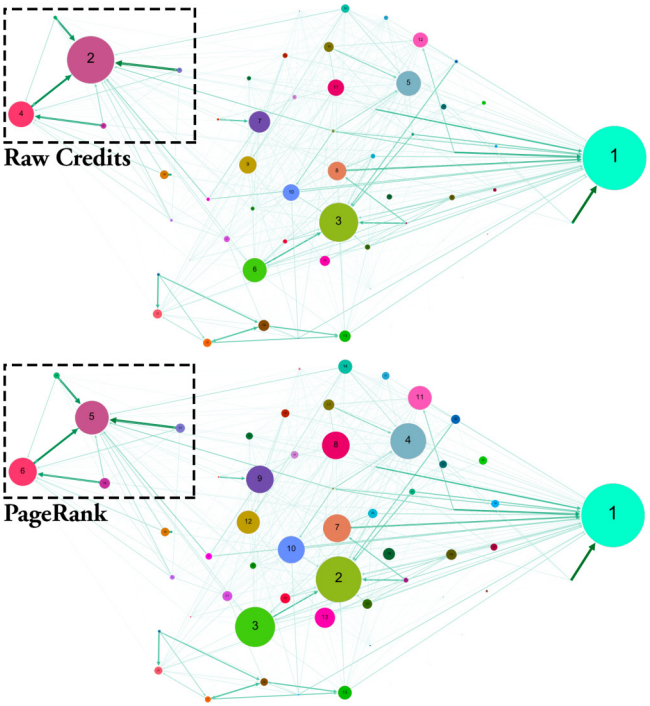

Figure 6. The credit network from a paper submission. Edge widths indicate the number of credits given, and node diameter represents the credit score. Top: raw credits, with a link ring (top left) directing credit inward. Bottom: PageRank-adjusted credits dampen the link ring, shifting the main beneficiary from 2nd to 5th author.

In our approach, we give each participant 100 credit points that they can privately allocate to other participants based on their assessment of who impacted the project. For example, participants might assign credit to those submitting strong milestones, collaborating actively, or DRI-ing. These credit allocations create a weighted directed graph, where each node is a participant and the edge weight is the number of credits that one participant assigned to another. Intuitively, the graph encodes the credit that participants grant to each other.

We then translate the decentralized credit graph into a credit score for each participant. There are many possible transformations: we use graph centrality via PageRank [55], because centrality captures the concept of a universally-recognized participant. Whereas PageRank propagates score equally across all outgoing links, we modify the algorithm to propagate scores in proportion to the outgoing edge weights to capture participants' exact credit distributions. Suppose that $G = (P, C)$ is the credit graph, where P is the set of participants and C is the set of directed weighted edges. Suppose further that $C(i, j) \in [0, 1]$ represents the proportion of credits that participant i gave to j, and d is the PageRank damping factor (typically $d = 0.85$). Then our modified PageRank score $\rho_i(t)$ for participant i each iteration t is given by:

$$\rho_i(t) = \frac{1-d}{|P|} + d \sum_{p \in P} (\rho_p(t-1) \cdot C(p,i))$$

These PageRank scores induce a ranking on participants.

Given these PageRank scores, the principal investigator and DRIs work together to set a threshold score for co-authorship, based on their assessment of the level of contribution appropriate to be listed as a coauthor. Those below the cutoff are credited in acknowledgments. The PI also uses these credit rankings as a quantitative measure in letters of recommendation sent in support of participants.

We overcome the quid-pro-quo attack and link rings by manipulating the PageRank damping factor d and by limiting the fraction of a node's score that it can pass to any individual outlink [2]. However, there remain several degrees of freedom in this credit system. First: is the PI included in the credit graph? Excluding is appealing but in practice led to situations where the PI had no power to help resolve credit infighting, so we now include the PI in the credit graph. Second: when can participants see the results? In order to prevent post-submission authorship surprises, we collect initial credit distributions one week before the paper deadline and publicly publish a set of tentative PageRank scores. We then allow participants to change their credit distributions until a few hours before the deadline. Late credit changes affected mainly the ranking of the last authors.

DEPLOYMENT

Evaluation of Crowd Research requires understanding (i) whether the crowdsourcing technique enabled the achievement of crowd-led research, (ii) whether the technique supported access for those without traditional avenues for doing research, and (iii) what impact the decentralized credit distribution technique had on contributors' rankings.

We have run three Crowd Research projects over two years. The projects enrolled 1697 participants from 62 countries and six continents, and produced crowd-authored papers at top-tier Computer Science venues including ACM UIST [22] and ACM CSCW [80]. Despite having a median of zero other letter writers from institutions ranked above 500 worldwide, Crowd Research participants have gone on to further study at undergraduate and graduate programs at universities such as Stanford, UC Berkeley, Carnegie Mellon University, and MIT.

Project case study summaries

Participants' highest or in-progress degree was 2% high school, 73% undergraduate, 22% master's, and 3% Ph.D. 28% of participants were women, though this number varied by project: the computer vision project was overwhelmingly male, but the HCI project was 47% women. The median age was 21. 71% reported an engineering area of study. Participants included not just students and researchers but also, e.g., a data scientist on Wall Street, an ITP-trained designer, a TR35 India winner, and several professional software engineers.

Computer vision: hybrid vision algorithms

The computer vision project was the first Crowd Research deployment, and had the least structure: in the initial weeks, the aim of the project was intentionally kept vague and open to exploration. Participants spread out to find and summarize recent computer vision papers, then began an iterative process of formulating project proposals based on the review. Peers and the PI reviewed these proposals weekly. Eventually the PI aligned everyone on one team's proposal for integrating human workers with black-box classifiers to optimize performance at a certain crowdsourcing dollar cost.

With this new focus, participants developed datasets and evaluation procedures. However, many participants grew discouraged because their proposals were not selected. Some teams stepped up their work and became more collaborative, but others became far less active. This observation led the HCI and data science projects to keep project ideation more collective and less parallel, avoiding the abrupt cutoff of all but a single idea. In the final phase, the different groups worked in parallel to build interfaces, implement machine classifiers and perform experiments. The group published a work-in-progress poster at HCOMP 2015 with 54 authors [79].

HCI: the Daemo crowdsourcing platform

The HCI project, which created a new paid crowdsourcing platform, spent its initial weeks needfinding by interviewing workers and requesters, then iteratively flared, focused, and rapidly prototyped research ideas. The crowd led ideation with feedback from the PI. As the research ideas solidified, participants self-selected which to participate in, each under different DRIs: the design and engineering of the platform — called Daemo — a new reputation system for Daemo, and an open governance structure for Daemo. These groups iterated on interaction design, engineering, user study design and analysis, and writing, again each under DRIs. After twelve weeks, the PI onboarded a second cohort of participants who joined the first group, continued work, and collectively published a work-in-progress poster at UIST 2015 with 70 authors [20].

The group continued to work and submitted an integrated Daemo paper to CHI with 50 authors, but it was rejected principally for covering too many research thrusts in one paper. The group onboarded a third cohort a few months later and split the CHI submission into multiple papers. They published a full paper on the Boomerang reputation system at UIST 2016 with 37 authors [22], and a full paper on Crowd Guilds at CSCW 2017 with 28 authors [80]. Daemo has launched in private beta [21], and a paper based on data its workers collected won the best dataset paper award at EMNLP 2016 [60]. The group continues work to launch Daemo publicly.

Data science: testing the wisdom of crowds at scale

The data science project began with a literature review. Each participant found and summarized papers about the wisdom of crowds, extracting metadata about the task, the sample size, and the aggregation method. This produced 144 unique papers. Participants then synthesized 190 domains (e.g., calorie estimation, sports game prediction), which they narrowed by popular vote and PI input down to 50. The PI next provided a question template that would allow the experiment to be deployed at scale. Participants curated 20 questions for each domain, including any images or audio clips, and committed them to a GitHub repository.

A small team of highly-motivated participants coded the experimental infrastructure to deploy these questions, and the group ran a pilot experiment. Participants analyzed the results from whatever angle seemed most interesting to them, and

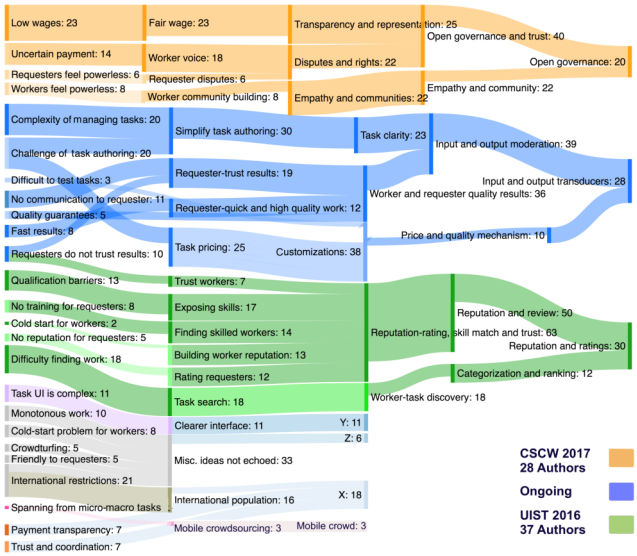

Figure 7. The HCI crowd produced and iterated ideas over seven weeks to develop three efforts, two of which are published papers now (orange and green). Darker shades were shared in the weekly meeting. Numbers report how many submissions proposed each idea.

submitted a short report outlining their findings. The group published a work-in-progress poster at UIST 2015 with 60 authors [66]. The final paper is currently in preparation.

The crowd led projects' ideation, execution, and writing

Where did the research insights come from? We inductively generated themes using the milestone submissions each week from the HCI project, labeled each submission with a theme, and coded related themes across weeks.

The final research directions trace back to the crowd's early brainstorms (Figure 7). These themes evolved into three ideas, two of which are now published. One set of themes in the first week, orange in Figure 7 (low wages, uncertain payment, feelings of powerlessness among workers, and feelings of powerlessness among requesters), evolved into Crowd Guilds at CSCW [80] and Daemo's open governance strategy. The green ideas in Figure 7 included a lack of trust in result quality, no training for requesters, and qualification barriers. These evolved into a redesign of reputation systems, producing Boomerang at UIST [22]. The crowd made other suggestions that the PI chose not to echo back to the crowd, lighter-shaded in Figure 7. Typically the PI did not echo ideas that were already played out in the research literature, for example the creation of a mobile crowdsourcing platform, or that did not constitute research goals, like addressing international restrictions for working on Mechanical Turk.

The crowd led paper writing as well (Figure 8). We analyzed the edit history of the shared text editor used for one of the HCI project papers. Each edit represents insertion or deletion of a block of text. The crowd made 8,360 edits (84%), while the principal investigator made 1,580 edits (16%). The PI focused their edits mainly on the sections that frame the paper, such as the Introduction (Figure 8). We compared this distribution to five papers in similar venues by the same PI but with their traditional Ph.D. students. On average, the Ph.D. students

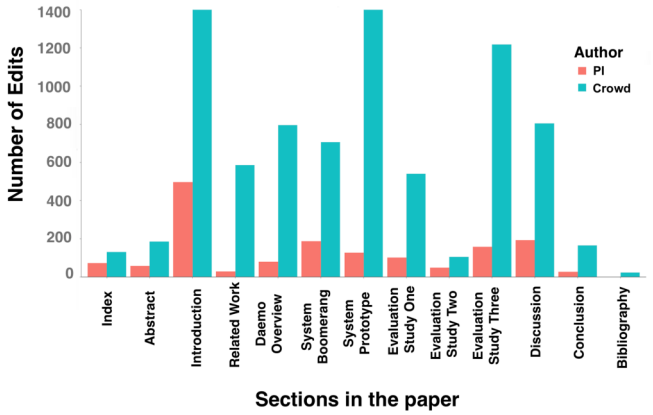

Figure 8. The crowd led paper writing. In this submission, the crowd made 84% of paper edits, and the PI 16%.

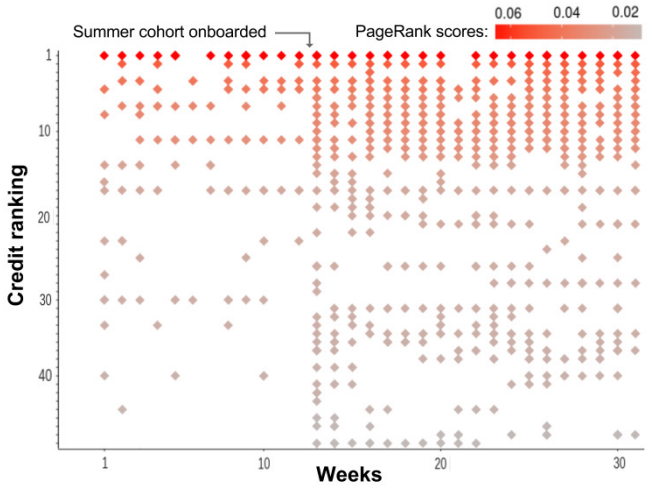

Figure 9. Each row of dots represents a participant's weeks of active participation. The y-axis is the author order position.

made 85% of the edits, and the PI 15% (σ=7%). So, this distribution is consistent with the PI's usual writing patterns.

Overall, Crowd Research enabled and empowered the crowd to choose domains of interest and lead diverse efforts. As one participant echoed via a survey: *"I really enjoyed the freedom to collaborate and try out different tasks. I initially thought I would be on the coding side, but I found myself leading my group on open gov[ernance] and design initiatives from which I was able to successfully communicate and learn."*

Participants remained active for months

For Crowd Research to be effective, its participants must stay dedicated for a long period of time: research does not happen overnight. We measured active participation via Slack activity, because it indicates ongoing investment in the effort: team milestone submissions allow hiding behind a single active member, but Slack participation is tagged to each individual.

After ten weeks, 15% of sign-ups and 29% of those who had participated in Slack were still active in the HCI project (Figure 9). Participants were occasionally inactive due to exams and life events. Across projects, crowd members exchanged

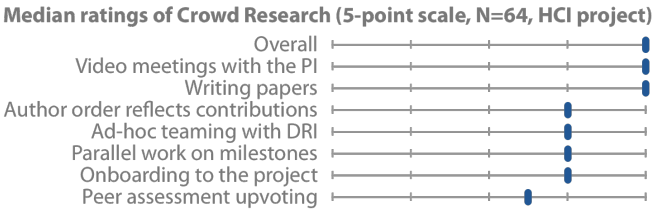

Figure 10. Participants reported authentic research experiences, useful coordination strategies, and accurate credit.

500,000 Slack messages (1,700 per week per project) and participated in 190,000 minutes of video meetings.

Several months after the projects launched, we surveyed active and inactive participants. Participants ($N = 173$) self reported a median 10 hours per week (mean 15hr), which is substantial on top of other courses. The most common self-reported reasons for dropout were the inability to catch up after exams (53% agreed), the level of time commitment (35%), and losing friends or teammates (17%). This ranking of reasons was consistent across the three projects.

For their part, the PIs' time commitment depended on each PI's advising style. Some focused only on group meetings, while others helped read submissions. In general, PIs spent two hours per week on the project: 1) a weekly one-hour meeting with RAs to understand progress and design milestones; 2) a weekly one-hour video advising meeting with the crowd; 3) sporadically helping with research-related questions over Slack. PIs engaged more heavily near deadlines.

Overall feedback was positive (Figure 10). One participant shared: *"This increases my interest in area of research. I learned many things from this project like writing research paper and skills like Angular JS & other frameworks, collaboration between team members and many more. It was a great enjoyable and educational learning experience."*

Crowd Research provided new routes for access

As evidence of access and upward mobility for traditional under-represented groups, we analyzed participants' self-reported age, gender, location, and affiliation from when they signed up. To understand whether participants had access to research experiences, we matched affiliations onto the Times Higher Education World University Rankings' subscale for research activity and influence [1], and locations at the country level onto a measure of GDP per capita [8].

Most participants did not have prior access to research experiences. 74% were at institutions ranked below 500 worldwide. 66% were in countries ranked below 50 in GDP per capita.

Crowd research papers were substantially more diverse, in terms of authors' affiliation and current country, than others in the same top-tier venues (Figure 11). We gathered all papers from CSCW 2017 and UIST 2016, where the papers were published, and compared the authors' affiliation rankings and country GDP per capita. The two Crowd Research papers had 57% and 58% of coauthors from universities ranked below 500 worldwide, vs. 12% and 11% of other papers in the venues (both $p < 0.001$). Likewise, the two Crowd Research papers had 42% and 35% of coauthors hailing from countries ranked

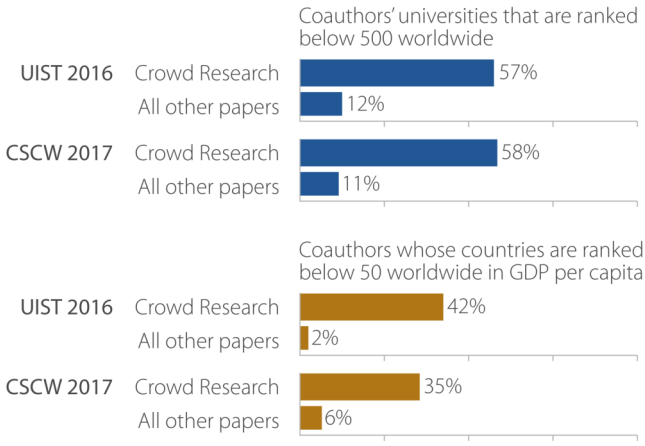

Figure 11. Crowd Research paper authors were more diverse than others at the conferences they appeared at. (all $p < 0.001$)

below the top 50 in GDP per capita, vs. 2% and 6% for others papers in the two venues (both $p < 0.001$).

Participants leveraged PIs' recommendation letters to gain access to education and jobs. We surveyed all participants who received a letter from a PI, and 33 responded. Of these, 21 received an offer from an institution or a company that they applied to. These participants also sought other letters; however, a median of 0 other letters (mean 0.37) were from universities or organizations that ranked above the top 500 worldwide. Thus, the Crowd Research PIs were the only recommenders from top-tier universities for many participants. These participants have since been admitted to undergraduate and graduate programs at schools including Stanford, UC Berkeley, Carnegie Mellon University, and MIT.

Feedback from participants emphasizes that they valued the access. As an undergraduate in India shared: *"It provided me with the opportunity to associate myself with top research work, and these opportunities weren't available to us back home. It also allowed me to learn about the research methodology as practiced in universities such as Stanford and definitely went a long way in helping me secure admission."*

Not everyone received admission — whether due to grades or insufficient contributions to Crowd Research to warrant a strong letter. They identified other benefits: *"While involvement with Crowd Research has given momentum to my pursuit of the future of problem solving and work, its effect on my current career as a librarian has been uncertain. I am not troubled by this though, because I believe any short term opportunity costs will be made up by long term benefits of having the foundation laid by my Crowd Research experience. [...] Any concern I have for my own career is far outweighed by my interest in shaping the nature of individual contributions to society at a large scale. That said, I have been able to secure funding for several conferences, am in the process of writing a white paper."*

Decentralized credit amplified concrete contributions

Participants felt that the author orders reflected their contributions (Median Likert 4/5, Figure 10). Figure 12 plots the

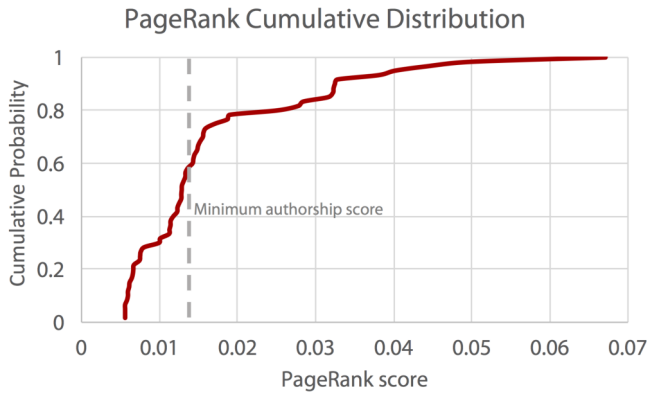

Figure 12. A CDF of the PageRank scores for the paper. Only those active up through the paper submission were eligible (after about one year). Of those, 36 were included as coauthors, and 41 were below the threshold and included in Acknowledgments instead.

PARTICIPATION MEASURE	PAGERANK SCORE β_{PR}	RAW SCORE β_{RS}	DIFFERENCE $\beta_{PR} - \beta_{RS}$
Meetings present	0.069^{***}	0.044^{*}	0.026
Files uploaded	0.035^{**}	0.029^{*}	0.006
GitHub commits	0.017	-0.024^{*}	0.041^{***}
Slack messages	0.035^{*}	0.112^{***}	-0.077^{***}
Self-organized meetings	0.024^{*}	0.012	0.012
Was DRI (binary)	0.036^{***}	0.006	0.030^{**}
Weeks active	0.025^{*}	0.014	0.011

Table 1. Regressions comparing the effects of participation behaviors on credit. Right column: compared to raw votes, PageRank increased the value of GitHub commits and DRI-ing, and decreased the value of talking on Slack. The median raw score was 0.06 (IQR [0.03, 0.15]), and the median distance between two adjacent authors was 0.003. The median PageRank score was 0.11 (IQR [0.03, 0.2]), and the median distance between two adjacent authors was 0.004. * p<.05; ** p<0.01; *** p<0.001.

cumulative distribution of PageRank scores for the CHI 2015 submission. There is a clustering of low scores representing about 60% of still-active accounts below the authorship threshold, and the remaining 40% more spread out amongst the higher scores.

But what effect did the networked credit allocation have on the author order? To investigate, we compared the initial raw credit scores and final PageRank-adjusted scores against logs of contributor behavior. We normalized the raw credit scores and the PageRank-adjusted scores to sum to 1.0. We then performed two multiple regressions, one predicting normalized raw score and one predicting PageRank score. Independent variables were observable participation behaviors, including Slack, GitHub, and weekly meetings, all standardized into z-scores. The regression coefficients β explain which behaviors were significantly correlated with changes in credit score. The right column of Table 1 highlights which features were significantly different between the raw credit and PageRank.

Relative to the raw score, PageRank lessened the impact of sending chat messages on Slack, and increased the impact of DRI-ing and committing code to GitHub (Table 1). This means that PageRank credit increased the effects of concrete contributions. These effects materially changed the author order. For example, one large team, who rarely interacted with the rest of the crowd, assigned nearly all of their credits to their team lead (Figure 6 top). Raw votes placed the team's lead second in the overall author order. However, PageRank softened the effect of this link ring (Figure 6 bottom), because others did not assign nearly as much credit to the team.

As one participant shared: *"It's obvious that if you have such a credit system, someone would try to cheat his/her reputation, but I think our credit system worked very well."* One point of frustration was last-minute contributors: *"Some people just appear two days before the paper submission deadline and take over the Slack channel and talk a lot. Then the result is that they are up-voted and got into the author list."* But most feedback was favorable: *"The idea of peer-evaluation and the use of PageRank resulted in mostly fair and accurate results."*

DISCUSSION

The Deployment section focused on successes. Here, we reflect on the challenges of Crowd Research. Challenges are instructive: they teach us the limitations of the technique, unpredicted outcomes, and opportunities for future research.

How to run a bad Crowd Research project

What lessons can be drawn for HCI and social computing? It can be more enlightening to discuss failure modes than successes. We offer a David Patterson-style list [58] of ways run a bad Crowd Research project:

1. *Assume 100% followthrough. Participants are extremely motivated, and work on Crowd Research to the exclusion of everything else.* Even motivated contributors have have jobs, exams, and lives. Milestones need to either utilize redundancy, or enforce deadlines and allow tasks to be reassigned if participants do not meet them.
2. *Encourage competition. Let the best contributions rise to the top.* This defaults to a critical culture, leading to dropout. It is critical to establish norms for a positive, inclusive culture [34]. In our case, switching from double-blind feedback to an upvoting system, plus consistent PI communication, helped changed the norm.
3. *Treat the crowd like incompetent undergraduates or mature graduate students.* Rote work leads to a lack of interest, but being too open-ended leaves many people behind. Balance the two through focused short-term milestones that encourage creativity.
4. *Pick projects you would do with your current lab.* This underplays the benefits of the crowd. It is better to leverage scale and diversity to achieve more ambitious goals.
5. *Assume that nobody will come into conflict.* Running a Crowd Research project feels like being in charge of a team or organization, giving rise to lots of progress but also interpersonal issues. This comes to a head especially around credit.

Limitations

One common question about Crowd Research is whether the effort is worth the PI's time investment of 2–3 hours per week. This is certainly higher than a single once-a-week meeting.

However, we would tend not to make a direct effort comparison. First, the three crowd research projects were more ambitious than typical projects in their respective labs, for example building a new crowdsourcing platform or running hundreds of experiments, making them difficult to compare to traditional papers. This was by design: we sought projects that capitalized on having a crowd. Second, the stated goal of Crowd Research is enabling access, not publishing more papers per hour. Empowering the crowd seemed worth an incremental extra time commitment.

A second critique is how much success can be attributed to the PI rather than the crowd. There was certainly PI-driven variation across projects: with Computer Vision the PI went on sabbatical and the project stopped after a WIP; with HCI the project was sustained through publication. However, the ideas themselves were crowd-driven (Figure 7), and like with advising traditional graduate students, most often the PI was helping filter bad ideas and amplify good ideas rather than propose all the ideas themselves.

A third question is to what extent any prestige associated with the universities were responsible for success. It seems likely that these names increased initial enrollment. However, the most popular project, Computer Vision, was the only one without a PI from a Top 15 university. This suggests that interest area rather than university name may have a substantial effect.

Disagreements and biases

Like any distributed team, conflicts broke out [29, 28]. Most commonly, these issues arose between participants: objections over the influence that someone was wielding, misaligned values (e.g., "talkers" vs. "doers"), disagreements on research decisions, or second-guessing of intentions. The PI or RAs diffused these situations, but they took an emotional toll on everyone involved. While most participants felt that author ordering was helpful, it was also a source of tension because votes were kept private. No matter how high someone was on the author list, we would hear complaints that they should have been ranked higher, or others ranked lower. In rare cases, participants publicly called each other out, which sowed tension and negatively affected trust.

A global project must also contend with cultural differences. The upside of cultural diversity is increased creativity and satisfaction [69, 40]. The downside is that diversity can lead to ethnocentrism, implicit and explicit biases [12]. Different cultures idealize different behaviors [77, 48]. Different cultures may also exhibit biases in how they treat women or other groups. If not carefully managed, cultural differences may drive out qualified participants or undervalue their contributions. It is nearly impossible to remove implicit biases from Crowd Research participants' credit evaluations of each other. Future work will measure the extent of these biases and identify ways to counteract them.

Future work

In the future, we hope to expand Crowd Research beyond Computer Science topics. In addition, we will make its suite of tools more turnkey so that any interested group can easily spin up a project. Finally, we hope to perform a longitudinal analysis or randomized trial to directly examine the long-term effects of participation.

CONCLUSION

This paper presents Crowd Research and an analysis of its two year long deployment. Crowd Research introduces a crowdsourcing technique for coordinating a large group of people in an open-ended research exploration, and a system for decentralized credit distribution. It enabled access to over 1,500 people worldwide to collaborate online in the pursuit of open-ended research. Utilizing Crowd Research, participants have built real-world systems, co-authored papers for top-tier conferences and have gone on to further careers in research.

Crowd Research represents a new form of knowledge production — one that leverages the diversity and scale of the internet to pursue projects that might be challenging in traditional laboratory environments. We believe that if Crowd Research and similar techniques successfully enable global access to training and mentorship experiences, they will help grow a new generation of scientists, humanists, and engineers that increase diversity in the scientific workforce. We envision that this generation could work collectively to resolve some of the biggest unanswered questions of our time.

ACKNOWLEDGEMENTS

We thank over 1,500 members of the Stanford Crowd Research Collective community for their contributions. This work was supported by Office of Naval Research awards N00014-16-1-2894 and N00014-15-1-2711, Institute for Scalable Scientific Data Management at UCSC and Los Alamos National Laboratory, Toyota, and the Hasso-Plattner Design Thinking Research Program.

REFERENCES

1. 2017. The Times Higher Education World University Rankings. (2017). https://www.timeshighereducation.com/world-university-rankings
2. Ricardo A Baeza-Yates, Carlos Castillo, Vicente López, Martin Shubik, John Hopcroft, and Daniel Sheldon. 2007. Pagerank Increase under Different Collusion Topologies. In *AIRWeb*, Vol. 5. Springer, Sage Publications, 68–81.
3. Julie A Bianchini. 2011. *Expanding underrepresented minority participation: America's science and technology talent at the crossroads*. National Academies Press, Washington, District of Columbia.
4. Rick Bonney, Heidi Ballard, Rebecca Jordan, Ellen McCallie, Tina Phillips, Jennifer Shirk, and Candie C Wilderman. 2009. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. (2009).
5. William G Bowen and Derek Bok. 2016. *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton University Press.
6. Nama Budhathoki. 2016. Who are the mappers and why do they map in OpenStreetMap. (2016). https://www.youtube.com/watch?v=LvakiUOsDrM

7. Yan Chen, Steve Oney, and Walter S Lasecki. 2016. Towards providing on-demand expert support for software developers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3192–3203.
8. CIA. 2016. The world factbook. (2016). `https://www.cia.gov/library/publications/the-world-factbook/`
9. Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015).
10. Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.
11. Joe Cox, Eun Young Oh, Brooke Simmons, Chris Lintott, Karen Masters, Anita Greenhill, Gary Graham, and Kate Holmes. 2015. Defining and measuring success in online citizen science: A case study of Zooniverse projects. *Computing in Science & Engineering* 17, 4 (2015), 28–41.
12. Catherine Durnell Cramton and Pamela J Hinds. 2004. Subgroup dynamics in internationally distributed teams: Ethnocentrism or cross-national learning? *Research in organizational behavior* 26 (2004), 231–263.
13. Justin Cranshaw and Aniket Kittur. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1865–1874.
14. Ruth Cronje, Spencer Rohlinger, Alycia Crall, and Greg Newman. 2011. Does Participation in Citizen Science Improve Scientific Literacy? A Study to Compare Assessment Methods. *Applied Environmental Education & Communication* 10, 3 (7 2011), 135–145.
15. Tawanna R Dillahunt, Bingxin Chen, and Stephanie Teasley. 2014. Model thinking: demographics and performance of MOOC students unable to afford a formal education. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 145–146.
16. Tawanna R Dillahunt, Sandy Ng, Michelle Fiesta, and Zengguang Wang. 2016. Do Massive Open Online Course Platforms Support Employability?. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM Press, New York, New York, USA, 232–243.
17. Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. 2009. PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology* 60, 11 (11 2009), 2229–2243.
18. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM Press, New York, New York, USA, 1013.
19. M Kevin Eagan, Sylvia Hurtado, Mitchell J Chang, Gina A Garcia, Felisha A Herrera, Juan C Garibay, and Juan C. Garibay. 2013. Making a Difference in Science Education: The Impact of Undergraduate Research Programs. *American Educational Research Journal* 50, 4 (8 2013), 683–713.
20. Snehalkumar (Neil) Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, Aditi Mithal, Adam Ginzberg, Aditi Nath, Karolina R. Ziulkoski, Trygve Cossette, Dilrukshi Gamage, Angela Richmond-Fuller, Ryo Suzuki, Jeerel Herrejón, Kevin Le, Claudia Flores-Saviaga, Haritha Thilakarathne, Kajal Gupta, William Dai, Ankita Sastry, Shirish Goyal, Thejan Rajapakshe, Niki Abolhassani, Angela Xie, Abigail Reyes, Surabhi Ingle, Verónica Jaramillo, Martin Godinez, Walter Angel, Carlos Toxtli, Juan Flores, Asmita Gupta, Vineet Sethia, Diana Padilla, Kristy Milland, Kristiono Setyadi, Nuwan Wajirasena, Muthitha Batagoda, Rolando Cruz, James Damon, Divya Nekkanti, Tejas Sarma, Mohamed Saleh, Gabriela Gongora-Svartzman, Soroosh Bateni, Gema Toledo Barrera, Alex Peña, Ryan Compton, Deen Aariff, Luis Palacios, Manuela Paula Ritter, Nisha K.K., Alan Kay, Jana Uhrmeister, Srivalli Nistala, Milad Esfahani, Elsa Bakiu, Christopher Diemert, Luca Matsumoto, Manik Singh, Krupa Patel, Ranjay Krishna, Geza Kovacs, Rajan Vaish, and Michael Bernstein. 2015. Daemo: A Self-Governed Crowdsourcing Marketplace. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 101–102.
21. Snehalkumar (Neil) S. Gaikwad, Mark E. Whiting, Dilrukshi Gamage, Catherine A. Mullings, Dinesh Majeti, Shirish Goyal, Aaron Gilbee, Nalin Chhibber, Adam Ginzberg, Angela Richmond-Fuller, Sekandar Matin, Vibhor Sehgal, Tejas Seshadri Sarma, Ahmed Nasser, Alipta Ballav, Jeff Regino, Sharon Zhou, Kamila Mananova, Preethi Srinivas, Karolina Ziulkoski, Dinesh Dhakal, Alexander Stolzoff, Senadhipathige S. Niranga, Mohamed Hashim Salih, Akshansh Sinha, Rajan Vaish, and Michael S. Bernstein. 2017. The Daemo Crowdsourcing Marketplace. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1–4.
22. S S Gaikwad, D Morina, A Ginzberg, C Mullings, S Goyal, D Gamage, C Diemert, M Burton, S Zhou, M Whiting, K Ziulkoski, A Ballav, A Gilbee, S S Niranga, V Sehgal, J Lin, L Kristianto, J Regino, N Chhibber, D Majeti, S Sharma, K Mananova, D Dhakal, W Dai, V Purynova, S Sandeep, V Chandrakanthan, T Sarma, S Matin, A Nassar, R Nistala, A Stolzoff, K Milland, V Mathur, R Vaish, and M S Bernstein. 2016. Boomerang: Rebounding the Consequences of Reputation Feedback

on Crowdsourcing Platforms. In *Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA.

23. C R Gerstner and D V Day. 1997. Meta-analytic review of leader-member exchange theory: Correlates and construct issues. *Journal of Applied Psychology* 82, 6 (1997), 827–844.
24. Eric Gilbert. 2013. Widespread underprovision on Reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 803–808.
25. Timothy Gowers and Michael Nielsen. 2009. Massively collaborative mathematics. *Nature* 461, 7266 (2009), 879–881.
26. Patricia Gurin, Eric Dey, Sylvia Hurtado, and Gerald Gurin. 2002. Diversity and higher education: Theory and impact on educational outcomes. *Harvard Educational Review* 72, 3 (2002), 330–367.
27. Russel S Hathaway, Biren A Nagda, and Sandra R Gregerman. 2002. The relationship of undergraduate research participation to graduate and professional education pursuit: an empirical study. *Journal of College Student Development* 43, 5 (2002), 614.
28. Pamela J Hinds and Diane E Bailey. 2003. Out of sight, out of sync: Understanding conflict in distributed teams. *Organization science* 14, 6 (2003), 615–632.
29. Pamela J Hinds and Mark Mortensen. 2005. Understanding conflict in geographically distributed teams: The moderating effects of shared identity, shared context, and spontaneous communication. *Organization science* 16, 3 (2005), 290–307.
30. Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.
31. Anne-Barrie Hunter, Sandra L. Laursen, and Elaine Seymour. 2007. Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education* 91, 1 (1 2007), 36–74.
32. Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 611–620.
33. Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, Mariusz Jaskolski, and David Baker. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology* 18, 10 (9 2011), 1175–1177.
34. Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* (2012).
35. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM Press, New York, New York, USA, 1301.
36. Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.
37. René F. Kizilcec and Sherif Halawa. 2015. Attrition and Achievement Gaps in Online Learning. In *Proceedings of the Second ACM Conference on Learning @ Scale*. ACM Press, New York, New York, USA, 57–66.
38. Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.
39. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 6 (2013), 33.
40. Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 75–84.
41. Karim R Lakhani, Hila Lifshitz-Assaf, and Michael Tushman. 2013. Open innovation and organizational boundaries: task decomposition, knowledge distribution and the locus of innovation. *Handbook of economic organization: Integrating economic and organizational theory* (2013), 355–382.
42. Kate Land, Anže Slosar, Chris Lintott, Dan Andreescu, Steven Bamford, Phil Murray, Robert Nichol, M Jordan Raddick, Kevin Schawinski, Alex Szalay, Daniel Thomas, and Jan Vandenberg. 2008. Galaxy Zoo: the large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 388, 4 (2008), 1686–1692.
43. R. Eric Landrum and Lisa R. Nelsen. 2002. The Undergraduate Research Assistantship: An Analysis of the Benefits. *Teaching of Psychology* 29, 1 (2 2002), 15–19.
44. Adam Lashinsky. 2012. *Inside Apple: How America's Most Admired–and Secretive–Company Really Works*. John Murray Publishers.
45. Thomas D LaToza, W Ben Towne, Christian M Adriano, and André Van Der Hoek. 2014. Microtask programming: Building software with a crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 43–54.

46. Edith Law, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. Crowdsourcing as a Tool for Research: Implications of Uncertainty. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '17)*. ACM, New York, NY, USA.

47. Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, and others. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6 (2014), 2122–2127.

48. Jin Li. 2003. US and Chinese cultural beliefs about learning. *Journal of educational psychology* 95, 2 (2003), 258.

49. David Lopatto. 2004. Survey of Undergraduate Research Experiences (SURE): first findings. *Cell biology education* 3, 4 (2004), 270–7.

50. Andrew Mao, Winter Mason, Siddharth Suri, Duncan J. Watts, and TW Malone. 2016. An Experimental Study of Team Size and Performance on a Complex Task. *PLOS ONE* 11, 4 (4 2016), e0153048.

51. David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM Press, New York, New York, USA, 224–235.

52. Robert C Miller, Haoqi Zhang, Eric Gilbert, and Elizabeth Gerber. 2014. Pair research: matching people for collaboration, learning, and productivity. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 1043–1048.

53. Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. 2016. WearWrite: Crowd-assisted writing from smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3834–3846.

54. Nigini Oliveira, Eunice Jun, and Katharina Reinecke. 2017. Citizen Science Opportunities in Volunteer-Based Online Experiments. (2017).

55. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web. (1999).

56. Scott E Page. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

57. Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. Gut Instinct: Creating scientific theories with online learners. (2017).

58. David A Patterson. 1994. How to have a bad career in research/academia. In *Keynote, 1994 USENIX Symposium on Operating System Design and Implementation*.

59. M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2009. Galaxy zoo: Exploring the motivations of citizen science volunteers. *arXiv preprint arXiv:0909.2925* (2009).

60. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

61. Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1364–1378.

62. Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 75–85.

63. Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. 2015. Best of both worlds: Human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2121–2131.

64. Susan H Russell, Mary P Hancock, and James McCullough. 2007. Benefits of undergraduate research experiences. *Science* 316, 5824 (2007), 548–549.

65. Heather Sarsons. 2015. Gender differences in recognition for group work. *Working Paper, Harvard University* (2015).

66. Alok Shankar Mysore, Vikas S Yaligar, Imanol Arrieta Ibarra, Camelia Simoiu, Sharad Goel, Ramesh Arvind, Chiraag Sumanth, Arvind Srikantan, Bhargav HS, Mayank Pahadia, Tushar Dobha, Atif Ahmed, Mani Shankar, Himani Agarwal, Rajat Agarwal, Sai Anirudh-Kondaveeti, Shashank Arun-Gokhale, Aayush Attri, Arpita Chandra, Yogitha Chilukur, 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 Maharaja, 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, Kasyap Varma-Dattada, Sukanya Venkataraman, Pulkit Verma, and Ishan Yelurwar. 2015. Investigating the "Wisdom of Crowds" at Scale. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15 Adjunct)*. ACM, New York, NY, USA, 75–76.

67. Hua-Wei Shen and Albert-László Barabási. 2014. Collective credit allocation in science. *Proceedings of the National Academy of Sciences* 111, 34 (2014), 12325–12330.

68. Jonathan Silvertown. 2009. A new dawn for citizen science. *Trends in Ecology & Evolution* 24, 9 (9 2009), 467–471.

69. Günter K Stahl, Martha L Maznevski, Andreas Voigt, and Karsten Jonsen. 2010. Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. *Journal of international business studies* 41, 4 (2010), 690–709.

70. Matthias Stevens, Michalis Vitos, Julia Altenbuchner, Gillian Conquest, Jerome Lewis, and Muki Haklay. 2014. Taking participatory citizen science to extremes. *IEEE Pervasive Computing* 13, 2 (2014), 20–29.

71. Brian L. Sullivan, Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (10 2009), 2282–2292.

72. Yizhou Sun, Yintao Yu, and Jiawei Han. 2009. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 797–806.

73. Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. 2016. Atelier: Repurposing Expert Crowdsourcing Tasks as Micro-internships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2645–2656.

74. Jaime Teevan, Shamsi T Iqbal, and Curtis von Veh. 2016. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2657–2668.

75. Ginka Toegel and Jay A Conger. 2003. 360-degree assessment: Time for reinvention. *Academy of Management Learning & Education* 2, 3 (2003), 297–311.

76. Bill Tomlinson, Joel Ross, Paul Andre, Eric Baumer, Donald Patterson, Joseph Corneli, Martin Mahaux, Syavash Nobarany, Marco Lazzari, Birgit Penzenstadler, and others. 2012. Massively distributed authorship of academic papers. In *Extended Abstracts of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, 11–20.

77. Jeanne L Tsai. 2007. Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science* 2, 3 (2007), 242–259.

78. Melissa A Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S Bernstein. 2017. Flash Organizations: Crowdsourcing Complex Work by Structuring Crowds As Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3523–3537.

79. Andreas Veit, Michael J Wilber, Rajan Vaish, Serge J Belongie, James Davis, Vishal Anand, Anshu Aviral, Prithvijit Chakrabarty, Yash Chandak, Sidharth Chaturvedi, Chinmaya Devaraj, Ankit Dhall, Utkarsh Dwivedi, Sanket Gupte, Sharath N Sridhar, Karthik Paga, Anuj Pahuja, Aditya Raisinghani, Ayush Sharma, Shweta Sharma, Darpana Sinha, Nisarg Thakkar, K Bala Vignesh, Utkarsh Verma, Kanniganti Abhishek, Amod Agrawal, Arya Aishwarya, Aurgho Bhattacharjee, Sarveshwaran Dhanasekar, Venkata Karthik Gullapalli, Shuchita Gupta, Chandana G, Kinjal Jain, Simran Kapur, Meghana Kasula, Shashi Kumar, Parth Kundaliya, Utkarsh Mathur, Alankrit Mishra, Aayush Mudgal, Aditya Nadimpalli, Munakala Sree Nihit, Akanksha Periwal, Ayush Sagar, Ayush Shah, Vikas Sharma, Yashovardhan Sharma, Faizal Siddiqui, Virender Singh, Abhinav S., Pradyumna Tambwekar, Rashida Taskin, Ankit Tripathi, and Anurag D Yadav. 2015. On Optimizing Human-Machine Task Assignments. *HCOMP 2015 Extended Abstracts* (2015).

80. Mark E Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, and Michael S Bernstein. 2017. Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1902–1913.

81. Haizi Yu, Biplab Deka, Jerry O Talton, and Ranjitha Kumar. 2016. Accounting for taste: ranking curators and content in social networks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2383–2389.

82. Lixiu Yu, Aniket Kittur, and Robert E Kraut. 2014. Distributed analogical idea generation: inventing with crowds. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1245–1254.

83. Haoqi Zhang, Matthew W. Easterday, Elizabeth M. Gerber, Daniel Rees Lewis, and Leesha Maliakal. 2017. Agile Research Studios. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 220–232.

84. Andrew L. Zydney, Joan S. Bennett, Abdus Shahid, and Karen W. Bauer. 2002. Impact of Undergraduate Research Experience in Engineering. *Journal of Engineering Education* 91, April (4 2002), 151–157.