BARCELONA 2013
Druid
ESADE Business School
Ramon Llull University

Paper to be presented at the

35th DRUID Celebration Conference 2013, Barcelona, Spain, June 17-19

Crowd Science: The Organization of Scientific Research in Open Collaborative Projects

Chiara Franzoni
Politecnico di Milano
DIG
chiara.franzoni@polimi.it

Henry Sauermann
Georgia Institute of Technology
Scheller College of Business
henry.sauermann@scheller.gatech.edu

Abstract

A growing amount of scientific research is done in an open collaborative fashion, in projects that are sometimes labeled as "crowd science", "citizen science", or "networked science". This paper seeks to gain a more systematic understanding of crowd science and to provide scholars with a conceptual framework and an agenda for future research. First, we briefly present three case examples that span different fields of science and illustrate the heterogeneity concerning what crowd science projects do and how they are organized. Second, we identify two fundamental elements that characterize crowd science projects - open participation and open sharing of intermediate inputs - and distinguish crowd science from other knowledge production regimes such as innovation contests or traditional "Mertonian" science. Third, we explore potential knowledge-related and motivational benefits that crowd science offers over alternative organizational modes, and potential challenges it is likely to face. Drawing on prior research on the organization of problem solving, we also consider for what kinds of tasks particular benefits or challenges are likely to be most pronounced. We conclude by outlining an agenda for future research and by discussing implications for funding agencies and policy makers.

Jelcodes:O31,O32

Crowd Science:

The Organization of Scientific Research in Open Collaborative Projects

Chiara Franzoni
DIG, Politecnico di Milano
P. Leonardo da Vinci 32, Milano, 20133
chiara.franzoni@polimi.it

Henry Sauermann
Georgia Institute of Technology, Scheller College of Business
800 W Peachtree St., Atlanta, GA 30308
henry.sauermann@scheller.gatech.edu

Abstract

A growing amount of scientific research is done in an open collaborative fashion, in projects that are sometimes labeled as "crowd science", "citizen science", or "networked science". This paper seeks to gain a more systematic understanding of crowd science and to provide scholars with a conceptual framework and an agenda for future research. First, we briefly present three case examples that span different fields of science and illustrate the heterogeneity concerning what crowd science projects do and how they are organized. Second, we identify two fundamental elements that characterize crowd science projects - open participation and open sharing of intermediate inputs - and distinguish crowd science from other knowledge production regimes such as innovation contests or traditional "Mertonian" science. Third, we explore potential knowledge-related and motivational benefits that crowd science offers over alternative organizational modes, and potential challenges it is likely to face. Drawing on prior research on the organization of problem solving, we also consider for what kinds of tasks particular benefits or challenges are likely to be most pronounced. We conclude by outlining an agenda for future research and by discussing implications for funding agencies and policy makers.

Keywords: crowd science; citizen science; crowdsourcing; community-based production; problem solving; open innovation; funding

April 8, 2013

We are grateful for stimulating discussions at the 10th International Open and User Innovation Workshop and the 2012 NBER Workshop on Scholarly Communication, Open Science, and its Impacts. We thank Pierre Azoulay, Rhiannon Crain, Lee Fleming, Lars Frederiksen, Carolin Haeussler, Lars Bo Jeppesen, Firas Khatib, Karim Lakhani, Danielle Li, Chris Lintott, Ali Mohammadi, Alex Oettl, Cristina Rossi-Lamastra, Arfon Smith, Ed Steinmueller, and Eric von Hippel for comments and suggestions. We also benefited greatly from feedback from two anonymous reviewers. All errors are our own.

1 Introduction

For the last century, scientific activity has been firmly placed in universities or other academic organizations, government laboratories, or in the R&D department of firms. Scholars in the sociology and economics of science, in turn, have made great progress understanding the functioning of this established system of science (Dasgupta & David, 1994; Merton, 1973; Stephan, 2012; Zuckerman, 1988). The last few years, however, have witnessed the emergence of projects that do not fit the mold of traditional science and that appear to follow distinct organizing principles. Foldit, for example, is a large-scale collaborative project involving thousands of participants who advance our understanding of protein folding at an unprecedented speed, using a computer game as their platform. Galaxy Zoo is a project involving over 250,000 volunteers who help with the collection of astronomical data, and who have contributed to the discovery of new classes of galaxies and a deeper understanding of the universe. Finally, Polymath involves a colorful mix of Fields Medalists and non-professional mathematicians who collectively solve problems that have long eluded the traditional approaches of mathematical science.

While a common term for these projects has yet to be found, they are variously described using labels such as "crowd science", "citizen science", "networked science", or "massively-collaborative science" (Nielsen, 2012; Wiggins & Crowston, 2011; Young, 2010). Even though there is significant heterogeneity across projects, they are largely characterized by two important features: participation in a project is open to a wide base of potential contributors, and intermediate inputs such as data or problem solving algorithms are made openly available. What we will call "crowd science" is attracting growing attention from the scientific community, but also policy makers, funding agencies and managers who seek to evaluate its potential benefits and challenges.[1] Based on the experiences of early crowd science projects, the opportunities are considerable. Among others, crowd science projects are able to draw on the effort and knowledge inputs provided by a large and diverse base of contributors, potentially expanding the range of scientific problems that can be addressed at relatively low cost, while also increasing the speed at which they can be solved. Indeed, crowd science projects have resulted in a number of high-profile publications in scientific outlets such as *Science*, PNAS, and *Nature Biotechnology*. At the same time, crowd science projects face important challenges in areas such as attracting contributors or coordinating the contributions of a large number of participants. A deeper understanding of these various benefits and challenges may allow us to assess the general prospects of crowd science, but also to

[1] For example, scientific journals have published special issues on citizen science, the topic has been discussed in managerial outlets such as the Sloan Management Review (Brokaw, 2011), national funding agencies in the US and other countries actively fund crowd science projects, and the Library of Congress is discussing how crowd science artifacts such as blogs and data sets should be preserved and curated.

conjecture for which kinds of scientific problems crowd science may be more - or less - suitable than alternative modes of knowledge production.

Despite the growing number of crowd science projects in a wide range of fields (see Table 1 for prominent examples), scholarly work on crowd science itself is largely absent. We address this lack of research in several ways. First, we introduce the reader to crowd science by briefly presenting three case studies of crowd science projects in biochemistry, astronomy, and mathematics. These case studies illustrate the heterogeneity concerning what crowd science projects do and how they are organized. Second, we identify organizational features that are common to crowd science projects while also distinguishing them from projects in "traditional science" and other emerging organizational paradigms such as crowd sourcing and innovation contests. Third, we discuss potential benefits and challenges crowd science projects are likely to face, how challenges may be addressed, and for what kinds of tasks the benefits and challenges may be most pronounced. In doing so, we build upon organizational theories of problem solving as well as prior work in areas such as open innovation and open source software development. We then outline an agenda for future research on crowd science and also discuss how crowd science may serve as a unique setting for the study of problem solving more generally. Finally, we discuss potential implications for funding agencies and policy makers.

2 Examples of crowd science projects

2.1 Foldit[2]

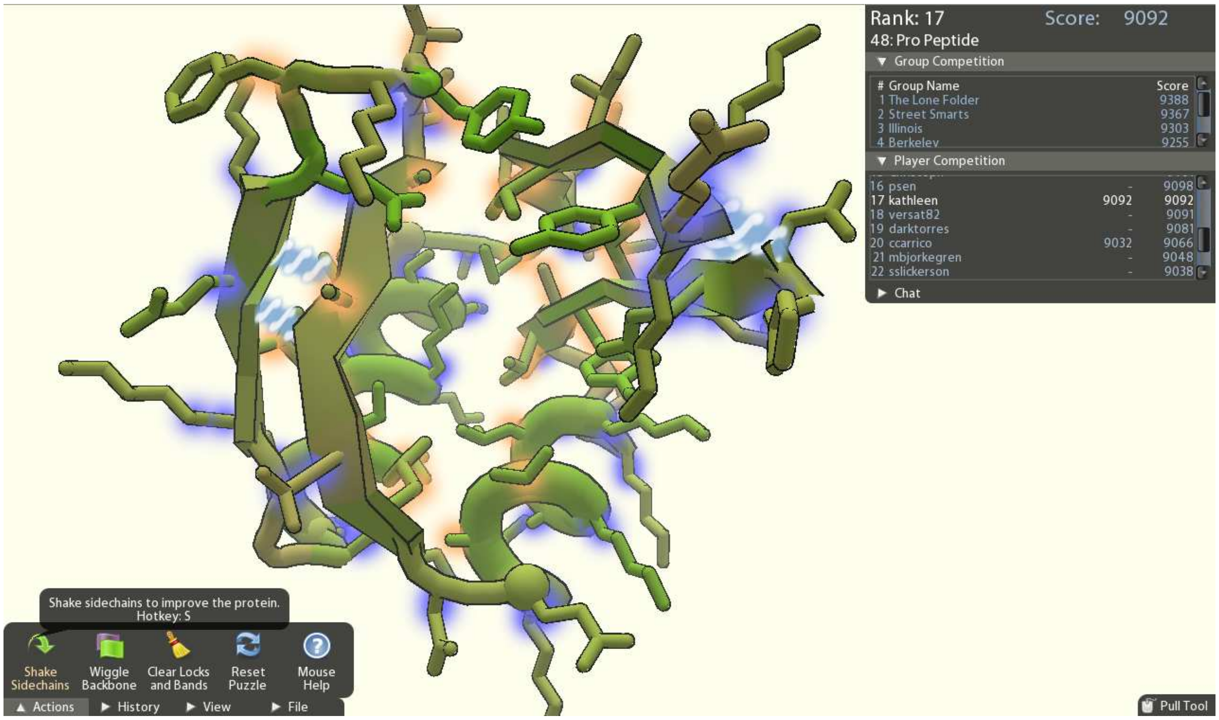
By the 1990s, scientists had developed significant insights into the biochemical composition of proteins. However, they had a very limited understanding of protein structure and shapes. Shape is important because it explains the way in which proteins function and interact with cells, viruses or proteins of the human body. For example, a suitably shaped protein could block the replication of a virus. Or it could stick to the active site of a biofuel and catalyze a chemical reaction. Conventional methods to determine protein shapes included X-ray crystallography, nuclear magnetic resonance spectroscopy, and electron microscopy. Unfortunately, these methods were extremely expensive, costing up to 100,000 dollars for a single protein, with millions of protein structures yet to be determined. In 2000 David Baker and his lab at the University of Washington, Seattle, received a grant from the Howard Hughes Medical Institute to work on shape determination with computational algorithms. Researchers believed that in principle, proteins should fold such that their shape employs the minimum level of energy to prevent the protein from falling apart. Thus, computational algorithms should be able to determine the shape of a

[2] The Foldit case study is based on the following web sources: http://fold.it; http://www.youtube.com/watch?v=2ad_ZW-mpOk; http://www.youtube.com/watch?v=PE0_48WhCCA; http://www.youtube.com/watch?v=nfxGnCcx9Ag; http://www.youtube.com/watch?v=uBA0vKURH3Y; retrieved September 16, 2012.

protein based on the electric charges of its components. However, each portion of a protein sequence is composed of multiple atoms and each atom has its own preferences for bonding with or standing apart from other atoms, resulting in a large number of degrees of freedom in a single molecule, making computational solutions extremely difficult. Baker and his lab developed an algorithm called Rosetta that combined deterministic and stochastic techniques to compute the level of energy of randomly chosen protein shapes in search of the best result. After several years of improvements, the algorithm worked reasonably well, especially on partially determined shapes. Because the computation was extremely intensive, in the fall of 2005 the team launched Rosetta@home, a grid system that allowed volunteers to make available the spare computational capacity of their personal computers. A critical feature of Rosetta@home was a visual interface that showed proteins as they folded. Although the volunteers were meant to contribute only computational power, looking at the Rosetta screensavers, some of them posted comments suggesting better ways to fold the proteins than what they saw the computer doing. These otherwise naïve comments inspired a group of post-docs at Baker's lab. They began to wonder if human visual ability could complement computer power in areas where the computer intelligence was falling short. Working with the department of computer science and engineering, they developed a web-based game called Foldit that enabled players to model the structure of proteins with the move of the mouse. Players could inspect a template structure from different angles. They could then move, rotate or flip chain branches in search of better structures. The software automatically computed the level of energy of new configurations, immediately showing improvement or worsening. Certain common structural problems, such as the existence of clashes or vacuums in the protein, were highlighted in red, so that the player could quickly identify areas of improvement. A menu of automatic tools borrowed from Rosetta enabled easy local adjustments that the computer could do better than humans. For example, proteins could be wiggled or side chains shaken with a simple click of the mouse. These features, combined with a few online tutorials, allowed people to start folding proteins without knowing virtually anything about biochemistry. Most interestingly, the software was designed as a computer game and included a scoreboard that listed players' performance. As such, players competed to climb up in the rankings and they could also set up teams and share strategies to compete against other teams. Figure 1 provides an impression of the Foldit interface.

The game was initially launched in May 2008. By September of the same year it had already engaged 50,000 users. The players were initially given known protein structures so that they could see the desired solution. After a few months of practice, several players had worked their way to shapes very close to the solution and in several cases had outperformed the best structures designed by Rosetta (Cooper et al., 2010). There was much excitement at the lab and the researchers invited a few top players to watch them play live. From these observations, it became clear that human intuition was very useful because it allowed players to move past the traps of local optima, which created considerable problems for computers. One year after launch there were about 200,000 active Foldit players.

Figure 1: Foldit user interface



Rank: 17 Score: 9092

48: Pro Peptide

▼ Group Competition

#	Group Name	Score
1	The Lone Folder	9388
2	Street Smarts	9367
3	Illinois	9303
4	Berkeley	9255

▼ Player Competition

#	Player		
16	psen	-	9098
17	kathleen	9092	9092
18	versat82	-	9091
19	darktorres	-	9081
20	ccarrico	9032	9066
21	mbjorkegren	-	9048
22	sslickerson	-	9038

Shake sidechains to improve the protein.
Hotkey: S

Source: http://fold.it/portal/info/about

In the following months, the development of Rosetta and that of Foldit proceeded in combination. Some proteins where Rosetta was failing were given to Foldit players to work on. In exchange, players suggested additional automatic tools that they thought the computer should provide for them. Meanwhile, players set up teams with names such as “Another hour, another point” or “Void Crushers” and used chats and forums to interact. Some players also began to elaborate their own “recipes”, encoded strategies that could be compared to those created in the lab. For example, some strategies involved wiggling a small part of a protein, rather than the entire structure, others were based on fusing protein parts or banding chains. Complex strategies are now often embedded in tools (“recipes”) that can subsequently be downloaded and used by others.[3]

The results were striking. A player strategy called “Bluefuse” completely displaced Rosetta “Classic Relax”, and outperformed “Fast Relax”, a piece of code that the Rosetta developers had worked on for quite a long time. These results were published in PNAS and the players of Foldit were co-authors under a collective pseudonym (Khatib et al., 2011a). In December 2010, encouraged by these results,

[3] http://foldit.wikia.com/wiki/Recipes Retrieved March 29, 2013.

Firas Khatib, Frank DiMaio, Seth Cooper and other post-docs working at Foldit in Baker's lab thought that the players were ready for a real-world challenge. Consulting with a group of experimentalists, they chose a monomeric retroviral protease, a protein known to be critical for monkey-virus HIV whose structure had puzzled crystallographers for over a decade. Two groups of players came to a fairly detailed solution of the protein structure in less than three weeks (published as Khatib et al., 2011b). As of September 2012, Foldit players were coauthors of four publications in top journals.

2.2 Galaxy Zoo[4]

In January 2006, a Stardust Spacecraft capsule landed in the Utah desert with precious samples of interstellar dust particles after having encountered the comet Wild 2. Particles in the sample were as tiny as a micron, and NASA took 1.6 million automated scanning microscope images of the entire surface of the collector. Because computers are not particularly good at image detection, NASA decided to upload the images online and to ask volunteers to visually inspect the material and report candidate dust particles. The experiment, known as Stardust@Home, had a large echo in the community of astronomers, where the inspection of large collections of images is a common problem.

In 2007, Kevin Schawinski, then a PhD in Chris Lintott's group at the University of Oxford, thought about using the same strategy, although the group's problem was different. They had hints that elliptical galaxies, contrary to the conventional theory, are not necessarily older than spiraling galaxies. In the spring of 2007 their insights were based on a limited sample of galaxies that Schawinski had coded manually, but more data were needed to really prove their point. A few months earlier, the Sloan Digital Sky Survey (SDSS) had made available 930,000 pictures of distant galaxies that could provide them with just the raw material they needed for their work. To be able to process this large amount of data, the researchers created the online platform Galaxy Zoo in the summer of 2007. Volunteers were asked to sign up, watch an online tutorial, and then code six different properties of astronomical objects visible in SDSS images. Participation became quickly viral, partly because the images were beautiful to look at, and partly because the BBC publicized the initiative on their blog. Before the project started, the largest published study was based on 9000 galaxies. Seven months after the project was launched, about 50 million galaxies had been coded and multiple classifications of a given galaxy by up to 50 different volunteers were used to reduce the incidence of incorrect coding. For an individual scientist, processing

[4] The Galaxy Zoo description is based on Nielsen (2012) and on the following web sources: http://www.galaxyzoo.org/story; http://supernova.galaxyzoo.org/; http://mergers.galaxyzoo.org/; http://www.youtube.com/watch?v=j_zQIQRr1Bo&playnext=1&list=PL5518A8D0F538C1CC&feature=results_main; http://data.galaxyzoo.org/; http://zoo2.galaxyzoo.org/; http://hubble.galaxyzoo.org/ http://supernova.galaxyzoo.org/about#supernovae; retrieved September 24, 2012.

50 million galaxies would have required more than 83 years of full-time effort.[5] The Galaxy Zoo data allowed Lintott's team to complete the study on the relationship between galaxy shape and age and to confirm their initial intuition that there is indeed a lot of new star formation in elliptical galaxies. However, this was just the beginning of Galaxy Zoo's contributions to science, partly because participants did not simply code galaxies - they also developed an increasing curiosity for what they saw. Consider the case of Annie van Arkel, a Dutch schoolteacher who in the summer of 2007 spotted an unusual "voorwerp" (Dutch for "thing") that appeared as an irregularly-shaped green cloud hanging below a galaxy. After her first observation, astronomers began to point powerful telescopes toward the cloud. It turned out that the cloud was a unique object that astronomers now explain as being an extinguished quasar whose light echo remains visible. Zooites also reported other unusual galaxies for their color or shape, such as very dense green galaxies that they named "Green pea galaxies" (Cardamone et al., 2009). A keyword search for "Annie's Voorwerp" in Web of Knowledge currently shows eight published papers, and a search for "Green pea galaxies" gives six.

The coded Galaxy Zoo data were made publicly available for further investigations in 2010. There are currently 141 scientific papers that quote the suggested citation for the data release, 36 of which are not coauthored by Lintott and his group.[6] After the success of the first Galaxy Zoo project, Galaxy Zoo 2 was launched in 2009 to look more closely at a subset of 250,000 galaxies. Galaxy Zoo Hubble was launched to classify images of galaxies made available by NASA's Hubble Space Telescope. Other projects looked at supernovae and at merging galaxies. Three years after Galaxy Zoo started, 250,000 Zooites had completed about 200 million classifications, and contributors are currently working on the latest and largest release of images from the SDSS.

The success of Galaxy Zoo sparked interest in various areas of science and the humanities. In 2010, Lintott and his team established a cooperation with other institutions in the UK and USA (the Citizen Science Alliance) to run a number of citizen science projects on a common platform "The Zooniverse", with financial support from the Alfred P. Sloan Foundation. The Zooniverse platform currently hosts projects in fields as diverse as astronomy, marine biology, climatology and medicine. Recent projects have also involved participants in a broader set of tasks and in closer interaction with machines. For example, contributors to the Galaxy Supernovae project were shown potentially interesting targets identified by computers at the Palomar Telescope. The contributors screened the large number of potential targets and selected a smaller subset that seemed particularly promising for further observation. This iterative process permitted astronomers to save precious observation time at large telescopes. Lintott

[5] Schawinski estimated his maximum inspection rate as being 50,000 coded galaxies per month. http://www.youtube.com/watch?v=j_zQIQRr1Bo&playnext=1&list=PL5518A8D0F538C1CC&feature=results_main; retrieved September 21, 2012.

[6] Search retrieved September 24, 2012.

thinks that in the future volunteers will be used to provide real-time judgments when computer predictions are unreliable, combining artificial and human intelligence in the most efficient way.[7]

2.3 Polymath[8]

Timothy Gowers is a British mathematician and a 1998 Fields Medal recipient for his work on combinatorics and functional analysis. An eclectic personality and an active advocate of openness in science, he keeps a regular blog that is well read by mathematicians. On January 29, 2009 he posted a comment on his blog saying that he would like to try an experiment to collectively solve a mathematical problem. In particular, he stated: "*The ideal outcome would be a solution of the problem with no single individual having to think all that hard. […] So try to resist the temptation to go away and think about something and come back with carefully polished thoughts: just give quick reactions to what you read […], explain briefly, but as precisely as you can, why you think it is feasible to answer the question*".[9] In the next few hours, several readers commented on his idea. They were generally in favor of trying the experiment and began to discuss practical issues like whether or not a blog was the ideal format for the project, and if the outcome should be a publication with a single collective name, or rather with a list of contributors. Encouraged by the positive feedback, Gowers posted the actual problem on February 1st: a combinatorial proof to the density version of the Hales-Jewett theorem. The discussion that followed spanned 6 weeks (see Figure 2 for an excerpt).

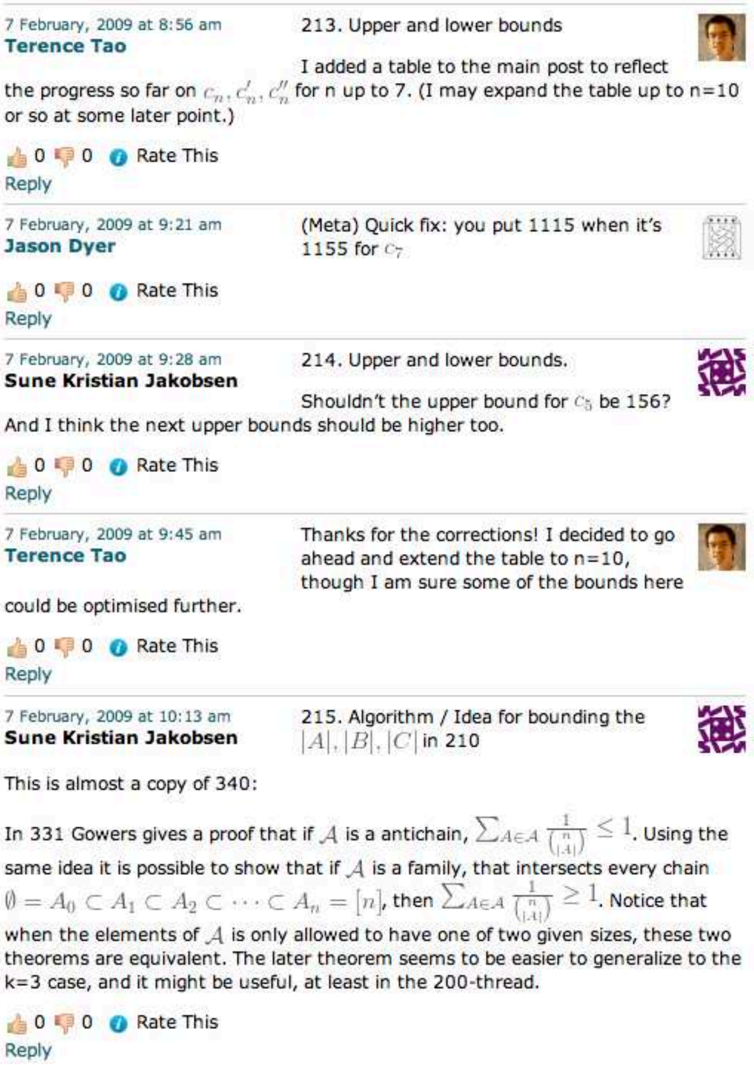
Among the contributors were several university professors, including Terry Tao, a top-notch mathematician at UCLA and also a Fields Medalist, as well as several school teachers and PhD students. After a few days, the discussion had branched out into several threads and a wiki was created to store arguments and ideas. Certain contributors were more active than others, but significant progress was coming form various sources. On March 10, Gowers announced that the problem was probably solved. He and a few colleagues took on the task of verifying the work and drafting a paper, and the article was eventually published in the Annals of Mathematics under the pseudonym "D.H.J. Polymath" (2012b).

[7] http://www.youtube.com/watch?v=j_zQIQRr1Bo&playnext=1&list=PL5518A8D0F538C1CC&feature=results_main; retrieved September 21, 2012.

[8] The Polymath case study is based on Nielsen (2012) as well as the following sources: http://gowers.wordpress.com/2009/01/27/is-massively-collaborative-mathematics-possible/ http://gowers.wordpress.com/2009/01/30/background-to-a-polymath-project/; http://gowers.wordpress.com/2009/02/01/a-combinatorial-approach-to-density-hales-jewett/; http://gowers.wordpress.com/2009/03/10/problem-solved-probably/; http://mathoverflow.net/questions/31482/the-sensitivity-of-2-colorings-of-the-d-dimensional-integer-lattice. http://gilkalai.wordpress.com/2009/07/17/the-polynomial-hirsch-conjecture-a-proposal-for-polymath3/; http://en.wordpress.com/tag/polymath4/; http://gowers.wordpress.com/2010/01/06/erdss-discrepancy-problem-as-a-forthcoming-polymath-project/; http://gowers.wordpress.com/2009/12/28/the-next-polymath-project-on-this-blog/; retrieved September, 2012.

[9] http://gowers.wordpress.com/2009/01/27/is-massively-collaborative-mathematics-possible/, retrieved September, 2012.

Figure 2: Excerpt from Polymath discussion



7 February, 2009 at 8:56 am
Terence Tao

213. Upper and lower bounds

I added a table to the main post to reflect the progress so far on c_n, c'_n, c''_n for n up to 7. (I may expand the table up to n=10 or so at some later point.)

👍 0 👎 0 Rate This
Reply

7 February, 2009 at 9:21 am
Jason Dyer

(Meta) Quick fix: you put 1115 when it's 1155 for c_7

👍 0 👎 0 Rate This
Reply

7 February, 2009 at 9:28 am
Sune Kristian Jakobsen

214. Upper and lower bounds.

Shouldn't the upper bound for c_5 be 156? And I think the next upper bounds should be higher too.

👍 0 👎 0 Rate This
Reply

7 February, 2009 at 9:45 am
Terence Tao

Thanks for the corrections! I decided to go ahead and extend the table to n=10, though I am sure some of the bounds here could be optimised further.

👍 0 👎 0 Rate This
Reply

7 February, 2009 at 10:13 am
Sune Kristian Jakobsen

215. Algorithm / Idea for bounding the $|A|, |B|, |C|$ in 210

This is almost a copy of 340:

In 331 Gowers gives a proof that if $\mathcal{A}$ is a antichain, $\sum_{A\in\mathcal{A}} \frac{1}{\binom{n}{|A|}} \leq 1$. Using the same idea it is possible to show that if $\mathcal{A}$ is a family, that intersects every chain $\emptyset = A_0 \subset A_1 \subset A_2 \subset \cdots \subset A_n = [n]$, then $\sum_{A\in\mathcal{A}} \frac{1}{\binom{n}{|A|}} \geq 1$. Notice that when the elements of $\mathcal{A}$ is only allowed to have one of two given sizes, these two theorems are equivalent. The later theorem seems to be easier to generalize to the k=3 case, and it might be useful, at least in the 200-thread.

👍 0 👎 0 Rate This
Reply

Source: http://terrytao.wordpress.com/2009/02/05/upper-and-lower-bounds-for-the-density-hales-jewett-problem/

Thrilled by the success of the original Polymath project, several of Gowers' colleagues launched similar projects, though with varying degrees of success. In June 2009, Terence Tao organized a collaborative entry to the International Mathematical Olympiads taking place annually in the summer. Similar projects have been successfully completed every year and are known as Mini-Polymath projects. Scott Aaronson began a project on the "sensitivity of 2-colorings of the d-dimensional integer lattice", which was active for over a year, but did not get to a final solution. Jil Kalai started a project on the "Polynomial Hirsh conjecture" (Polymath 3), again with inconclusive results. Terence Tao launched a project for finding primes deterministically (Polymath4), which was successfully completed and led to a collective publication under the pseudonym of D.H.J. Polymath in Mathematics of Computation (Polymath, 2012a). In January 2010 Timothy Gowers and Jil Kalai began coordinating a new Polymath project on the Erdos Discrepancy Problem (known as Polymath 5). An interesting aspect of this project is that the particular problem to be solved was chosen through a public discussion on Gowers' blog. Several Polymath projects are currently running. Despite the growing number of projects, however, the number of contributors to each particular project remains relatively small, typically not exceeding a few dozen.[10]

It is interesting to note that over time, Polymath projects have developed organizational practices that facilitate collective problem solving. In particular, a common challenge is that when the discussion develops into hundreds of comments, it becomes difficult for contributors to understand which tracks are promising, which have been abandoned and where the discussion really stands. The chronological order of comments is not always informative because people respond to different posts and the problems branch out in parallel conversations, making it difficult to continue discussions in a meaningful way. To overcome these problems, Gowers, Tao and other project leaders started to take on the role of moderators. When the comments on a topic become too many or too unfocused, they synthesize the latest results in a new post or open separate threads. Polymath also inspired mathoverflow.net, a useful tool for the broader scientific community. Launched in the fall of 2009, this platform allows mathematicians to post questions or problems and to provide answers or rate others' answers in a commented blog-style discussion. In some cases, the discussion develops in ways similar to a small Polymath project and answers have also been cited in scholarly articles.

2.4 Overview of additional crowd science projects

Table 1 shows additional examples of crowd science projects. The table specifies the primary scientific field of a project, illustrating the breadth of applications across fields such as astronomy, biochemistry, mathematics, or archeology. We also indicate what kind of task is performed by the crowd, e.g., the classification of images and sounds, the collection of observational data, or collective problem solving. This column shows the variety of tasks that can be accomplished in crowd science projects. We will draw on the earlier cases and the examples listed in Table 1 throughout our subsequent discussion.

----- Insert Table 1 here -----

3 Characterizing crowd science and exploring heterogeneity across projects

Examples such as those discussed in the prior section provide fascinating insights into an emerging way of doing science and have intrigued many observers. However, while there is agreement that these projects are somehow "different" from traditional science, a systematic understanding of the concrete nature of these differences is lacking. Similarly, it seems important to consider the extent to

[10] http://michaelnielsen.org/blog/the-polymath-project-scope-of-participation/ Retrieved September 28, 2012.

which these projects differ from other emerging approaches to producing knowledge such as crowd sourcing or innovation contests.

In the following section 3.1, we identify two key features that distinguish crowd science from other regimes of knowledge production: openness in project participation and openness with respect to the disclosure of intermediate inputs such as data or problem solving approaches. Not all projects share these features to the same degree; however, these dimensions tend to distinguish crowd science from other organizational forms, while also having important implications for opportunities and challenges crowd science projects may face. In the subsequent section 3.2, we will delve more deeply into heterogeneity among crowd science projects themselves, reinforcing the notion that "crowd science" is not a well-defined type of project but rather an emerging organizational mode of doing science that allows for significant experimentation, as well as considerable scope in the types of problems that can be addressed and in the types of people who can participate.

3.1 Putting crowd science in context: Different degrees of openness

A first important feature of crowd science that marks a difference to traditional science is that participation in projects is open to a large number of potential contributors that are typically unknown to each other or to project organizers at the beginning of a project. Any individual who is interested in a project is invited to join. Recall that Fields Medalist Timothy Gowers' invitation to join Polymath was accepted, among the others, by Terence Tao, another Fields Medalist working at UCLA, as well as a number of other less famous colleagues, schoolteachers, and graduate students. Open participation is even more salient in Galaxy Zoo, which recruits participants on its website and boasts a contributor base of over 250,000. Note that the emphasis here is not simply on a large number of project participants (large team size is becoming increasingly common even in traditional science, see Wuchty et al. (2007)). Rather, open participation entails virtually unrestricted "entry" by any interested and qualified individual, often based on self-selection in response to a general call for participation.

A second feature that tends to be common is that crowd science projects openly disclose a substantial part of the intermediate inputs used in the knowledge production such as data sets or problem solving approaches. The Whale Song Project, for example, has uploaded a database of audio recordings of whale songs, and Galaxy Zoo has made publicly available the classifications made by volunteer contributors (Lintott et al., 2010). As noted earlier, many Foldit players also share their tricks to facilitate certain operations, and complicated strategies are encoded in downloadable scripts and recipes. Thus, much of the process-related knowledge that has traditionally remained tacit in scientific research (Stephan, 1996) is codified and openly shared within and outside a particular project. We also subsume under the term "intermediate inputs" records of the actual process through which scientists develop and evaluate problem solving strategies since the resulting insights can facilitate future problem solving. Consider, for example, the following discussion of a problem solving approach from the Polymath 4 blog:

> *Anonymous (August 9, 2009 at 5:48pm) Tim's ideas on sumsets of logs got me to thinking about a related, but different approach to these sorts of "spacing problems": say we want to show that no interval [n, n + log n] contains only (log n)^100 – smooth numbers. If such strange intervals exist, then perhaps one can show that there are loads of distinct primes p in [(log n)^10, (log n)^100] that each divide some number in this interval […]*
>
> *Terrence Tao (August 9, 2009 at 6:00pm) I quite like this approach – it uses the entire set S of sums of the 1/p_i, rather than a finite sumset of the log-integers or log-primes, and so should in principle get the maximal boost from additive combinatorial methods. It's also using the specific properties of the integers more intimately than the logarithmic approach, which is perhaps a promising sign.*[11]

We suggest that these two dimensions – openness in participation and openness with respect to intermediate inputs – do not only describe common features of crowd science projects but that they also differentiate crowd science projects from other types of knowledge production.[12] In particular, Figure 3 shows how differences along these two dimensions allow us to distinguish in an abstract way crowd science from three other knowledge production regimes.

In Figure 3, crowd science contrasts most strongly with the bottom-left quadrant, which captures projects that limit contributions to a relatively small and pre-defined set of individuals and do not openly disclose intermediate inputs. We call this quadrant "traditional science" because it captures key features of the way science has been done over the last century. Of course, traditional science is often called "open" science because *final results* are openly disclosed in the form of publications (David, 2008; Murray & O'Mahony, 2007; Sauermann & Stephan, 2013). However, while traditional science is "open" in that sense, it is largely closed with respect to the dimensions captured in our framework. In particular, researchers retain exclusive use of their key intermediate inputs, such as the data and the heuristics used to solve problems. While some of these inputs are disclosed in methodology sections of papers, most of them remain tacit or are shared only among the members of the research team (Stephan, 1996).

This lack of openness follows from the logic of the reward system of the traditional institution of science. As emphasized by Merton in his classic analysis, traditional science places one key goal above

[11] http://polymathprojects.org/2009/08/09/research-thread-ii-deterministic-way-to-find-primes/ Retrieved 29 March 2012.
[12] Open participation and open disclosure of inputs is also characteristic of open source software (OSS) development, although the goal of the latter is not the production of scientific knowledge but of software artifacts. Given these similarities, our discussion will take advantage of prior work that has examined the organization of knowledge production in OSS development.

all others: gaining recognition in the community of peers by being the first to present or publish new research results (Merton, 1973; Stephan, 2012). While scientists may also care about other goals, publishing and the resulting peer recognition are critical because more publications translate into job security (tenure), more resources for research, more grants, more students, and so on. Since most of the recognition goes to the person who is first in discovering and publishing new knowledge, the institution of science induces scientists to expend great effort and to disclose research results as quickly as possible. At the same time, this system encourages scientists to seek ways to monopolize their area of expertise and to build a competitive advantage over rival teams. As such, the traditional reward system of science discourages scientists from helping contenders, explaining why data, heuristics, and problem solving strategies are often kept secret (Dasgupta & David, 1994; Haeussler et al., 2009; Walsh et al., 2005).

Figure 3: Knowledge production regimes with different degrees of openness

Project Participation	Disclosure of Intermediate Inputs: closed	Disclosure of Intermediate Inputs: open
open	Innovation Contests Crowd Sourcing (e.g., Innocentive, Mechanical Turk)	Crowd Science
closed	Traditional "Mertonian" Science	Traditional Science with Disclosure of Data and Logs (e.g., as required by funding agencies, journals)

Recognizing that secrecy with respect to intermediate inputs may slow down the progress of science, funding agencies increasingly ask scientists to openly disclose data and logs as a condition of funding. The National Institutes of Health and the Wellcome Trust, for example, require that data from genetic studies be made openly available. Similarly, more and more journals including the flagship publications *Science* and *Nature* require scientists to publicly deposit data and materials such that any interested scientist can replicate or build upon prior research.[13] As such, an increasing number of projects

[13] http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail; http://www.nature.com/authors/policies/availability.html; retrieved October 20, 2011.

have moved from quadrant 4 – "traditional science" – into quadrant 3 (bottom right). Even though projects in this quadrant are more open by disclosing intermediate inputs after the project is finished, participation in a given project remains closed to outsiders.

Let us finally turn to projects in quadrant 1 (top-left), which solicit contributions from a wide range of participants but do not publicly disclose intermediate inputs. Moreover, while not explicitly reflected in Figure 3, projects in this cell differ from the other three cells in that even final project results are typically not openly disclosed. Examples of this organizing mode include Amazon's Mechanical Turk (which pays individuals for performing tasks such as collecting data from websites) and – more interestingly – innovation contest platforms such as Kaggle or Innocentive. In innovation contests, "seekers" post their problems online and offer monetary prizes for the best solutions. Anybody is free to compete by submitting solutions, although many contests require that "solvers" first accept confidentiality and intellectual property agreements. Once a winner is determined, he or she will typically transfer the property right of the solution to the seeker in return for the prize (Jeppesen & Lakhani, 2010). Thus, the projects in quadrant 1 are open to a large pool of potential contributors, but they do not disclose results and data. A main reason for the latter is that project sponsors are often private organizations that seek to gain some sort of a competitive advantage by maintaining unique access to research results and new technologies. Along with crowd science, quadrant 1 is arguably the most dynamic, witnessing the rapid entry of new platforms and significant experimentation with different organizational designs. As such, our classification focuses on broad differences between quadrants and on common characteristics of examples as of the writing of this article.[14]

While openness with respect to project participation and with respect to intermediate inputs are by no means the only interesting aspects of crowd science projects, and while not all projects reflect these features to the same degree, these two dimensions highlight prominent qualitative differences between crowd science and other regimes of knowledge production. Moreover, we focus on these two dimensions because they have important implications for our understanding of potential benefits and challenges of crowd science. As highlighted by prior work on the organization of problem solving, the first dimension – open participation – is important because it speaks to the labor inputs and knowledge a project can draw on, and thus to its ability to solve problems and to generate new knowledge.[15] Openness with respect to

[14] There are interesting nuances in the designs chosen by the various players in quadrant 1, new players may enter with somewhat different organizational designs, and existing players may experiment with respect to their degrees of openness. A deeper understanding of the implications of crowd science's high degrees of openness may be fruitful in future work examining heterogeneity and changes in the degree of openness of projects currently located in other cells of Figure 3.

[15] The literature on open innovation and the governance of problem-solving activities in organizations focuses on the decision to locate problem solving within a given organizational boundary ("closed") versus involving agents located outside the organization ("open") (Afuah & Tucci, 2012; Felin & Zenger, 2012; Jeppesen & Lakhani, 2010; Nickerson & Zenger, 2004). While this prior work centers on firm boundaries, many of the resulting insights are useful in our context if we conceptualize the relevant organizational unit not as the firm but as the "core" team of researchers that would work on a problem in traditional science but now has the option to invite participation by the larger crowd.

intermediate inputs – our second dimension – is an important requirement for distributed work by a large number of crowd participants. At the same time, our discussion of the role of secrecy in traditional science suggests that this dimension may have fundamental implications for the kinds of rewards project contributors are able to appropriate and thus for the kinds of motives and incentives projects can rely on in attracting participants. We provide a more detailed discussion of the benefits and challenges resulting from open participation and open disclosure of intermediate inputs in sections 4 and 5.

3.2 Heterogeneity within: Differences in the nature of the task and in contributor skills

Section 3.1 highlighted two important common features of crowd science projects. However, there is also considerable heterogeneity among projects. In seeking deeper insights into this heterogeneity, we focus on two dimensions that may have particularly important implications for the benefits and challenges crowd science projects face. As reflected in Figure 4, these dimensions include the nature of the task that is performed by the crowd (horizontal axis), and the skills that projects participants need in order to make meaningful contributions (vertical axis).[16] We discuss each dimension in turn.

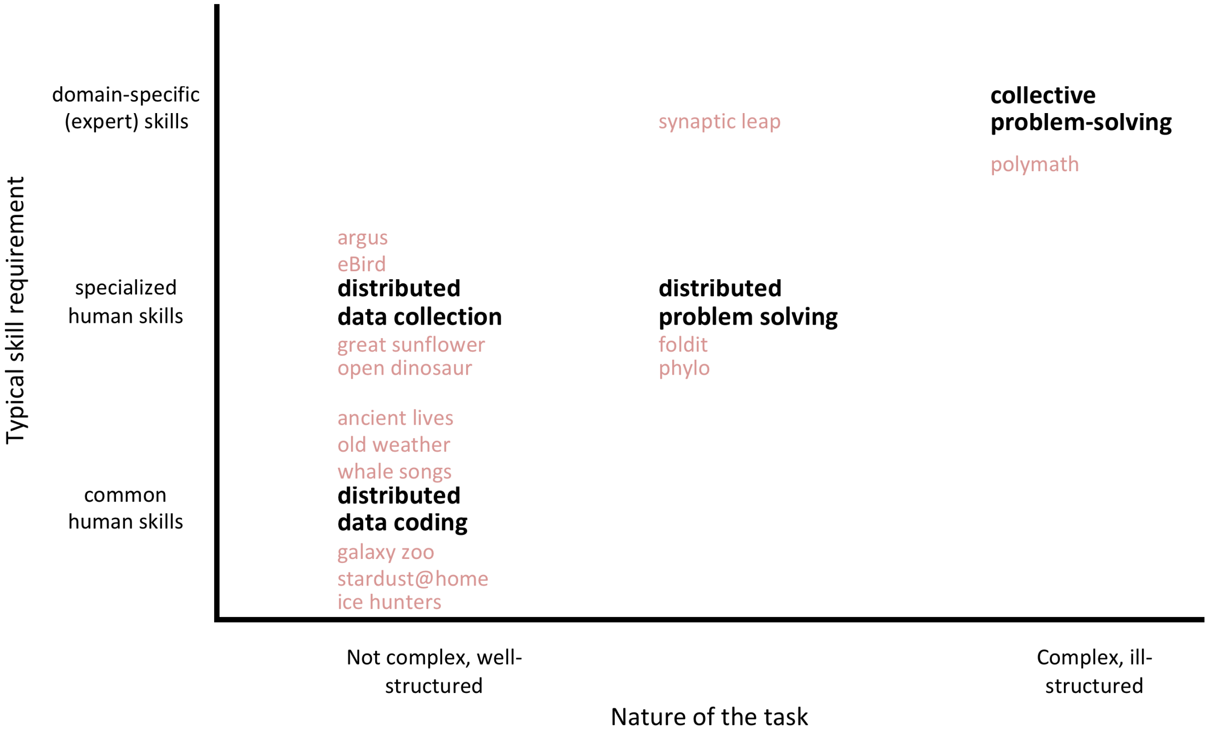
In a general sense, all crowd science projects ultimately have the goal to gain a deeper understanding of the natural world and to find solutions to scientific problems. As part of a project, the crowd performs certain types of tasks such as the coding of astronomical data, the optimization of protein structures, or the development of complex mathematical proofs. Note that we conceptualize the "task" in an aggregate sense at the level of the crowd, with each individual making distinct contributions to that task by engaging in certain subtasks. The tasks performed by the crowd differ in many ways, but prior research on the organization of problem solving suggests that it is particularly useful to consider differences in two related characteristics: the complexity of the task and the task structure (Felin & Zenger, 2012; Simon, 1962, 1973).

In our context, *task complexity* is best conceptualized as the degree of interdependency between individual subtasks. In tasks that are not complex, the best solution to a subtask is independent of other subtasks, allowing contributors to work independently. For example, the correct coding of an image in Galaxy Zoo does not depend on the coding of other images. In complex tasks, however, the best solution to a subtask depends on other subtasks and, as such, a contributor needs to consider other contributions when working on his/her own subtask. The term *task structure* captures the degree to which the task is "well-structured" versus "ill-structured". Well-structured tasks involve a clearly defined set of subtasks, the criteria for evaluating contributions are well understood, and the "problem space" is essentially mapped out in advance (Simon, 1973). In ill-structured tasks, the specific subtasks that need to be

[16] As we will discuss below, most projects have dedicated project organizers who often initiate the projects and closely resemble the principal investigators in traditional scientific projects. However, our focus here is on the typically much larger group of contributors outside of the immediate "core" of project organizers.

performed are not clear ex ante, contributions are not easily evaluated, and the problem space becomes clearer as the work progresses and as contributions build on each other. While task complexity and structure are distinct concepts, they are often related in that complex tasks tend to also be more ill-structured, partly due to our limited understanding of the interactions among different subtasks (Felin & Zenger, 2012). Considering differences across projects with respect to the nature of the task is particularly important because complex and ill-structured tasks provide fewer opportunities for the division of labor and may face distinct organizational challenges. We will discuss these challenges as well as potential mechanisms to address them in section 5 below.

Figure 4: Heterogeneity among crowd science projects



As Figure 4 illustrates, a large share of crowd science projects involves primarily tasks of low complexity that tend to be well-defined. For example, the coding of images for the Galaxy Zoo project involves a large number of individual contributions, but subtasks are independent and contributors can work in parallel without regard to what other contributors are doing.[17] At the other extreme are projects

[17] One may think that these coding tasks are so simple that their performance should not be considered as a serious scientific contribution at all. However, data collection is an integral part of scientific research and much of the effort (and money) in traditional research projects is expended on such data collection efforts, often carried out by PhDs and Postdocs. In fact, recall that Galaxy Zoo was initiated by a PhD student who was overwhelmed by the task to code a large number of galaxies. Reflecting

that involve highly complex and ill-defined tasks requiring contributors to build on each others' work in a sequential fashion, and to develop a collective understanding of the problem space and of possible solutions over time. Polymath projects exhibit these characteristics and Figure 2 illustrates the interactive problem solving process. Finally, projects located in the middle of the horizontal axis in Figure 4 involve tasks that are of moderate complexity and relatively well-defined, while still allowing contributors to collaborate and to explore different approaches to solve the problem. For example, Foldit players can find the best protein structure using very different approaches that are often discovered only in the process of problem solving. Moreover, while some players work independently, many of the most successful solutions have been found by members of larger teams who developed solutions collaboratively by building on each other's ideas.

Let us know consider the vertical axis in Figure 4, which reflects the different levels of skills that are required to make meaningful contributions to a project (Erickson et al., 2012).[18] As noted earlier, many "citizen science" projects such as Galaxy Zoo ask for contributions that require only skills that are common in the general human population. Other projects, such as Polymath, typically require expert skills in specific scientific domains, as evidenced by the prevalence of trained mathematicians among Polymath participants, or chemists and biomedical scientists in the Synaptic Leap project. Projects located in the middle of the vertical axis require skills that are specialized and less common but that are not tied to a particular scientific domain. For example, the Argus project relies on captains of ships to collect measures of seabed depth using sonar equipment, requiring specialized nautical skills. Similarly, success in protein folding requires the ability to visualize and manipulate three-dimensional shapes in space, a special cognitive skill that has little to do with the traditional study of biochemistry.[19]

4 Potential benefits of crowd science

In section 3.1, we compared crowd science to other knowledge production regimes and suggested that crowd science projects tend to be more open with respect to participation as well as the disclosure of intermediate inputs. We now discuss how this high degree of openness provides certain advantages with respect to the creation of new knowledge as well as the motivation of project participants. Section 5 considers potential challenges openness may pose and how these challenges may be addressed. Throughout these discussions, we will consider how the benefits and challenges may differ across different types of projects, focusing primarily on the distinctions drawn in the prior section 3.2.

this important function, contributions in the form of data collection or the provision of data are often explicitly rewarded with co-authorship on scientific papers (Haeussler & Sauermann, 2013).

[18] We focus on the kinds of skills required to make a meaningful contribution to a project and abstract from the significant heterogeneity in skills among contributors to a given project. In particular, it is likely that some contributors benefit from higher levels of skills or access to a broader range of skills than others.

[19] It is interesting to note that Foldit effectively uses sophisticated software to transform a task that traditionally required scientific training into a task that can be performed by non-scientists using a different set of skills.

4.1 Knowledge-related benefits

The scholarly literature has developed several different conceptualizations of the science and innovation process, emphasizing different problem solving strategies such as the recombination of prior knowledge, the search for extreme value outcomes, or the systematic testing of competing hypotheses (see Afuah & Tucci, 2012; Boudreau et al., 2011; Fleming, 2001; Kuhn, 1962; Weisberg, 2006). Given the wide spectrum of problems crowd science projects seek to solve, it is useful to draw on multiple conceptualizations of problem solving in thinking about the benefits projects may derive from openness in participation and from the open disclosure of intermediate results.

4.1.1 Benefits from open participation

Some projects such as Galaxy Zoo or Old Weather benefit primarily from a larger quantity of labor inputs, using thousands of volunteers to complement scarce material resources such as telescopes but also the unique expertise of highly trained project leaders. Could projects rely on powerful computers to accomplish these tasks without the involvement of a larger number of people? It turns out that humans continue to be more effective than computers in several realms of problem solving, including image and sound detection (Malone & Klein, 2007). Humans also rely on intuition to improve optimization processes, whereas computer algorithms often rely on random trial-and-error search and may get trapped in local optima. In particular, humans have a superior capacity to narrow the space of solutions, which is useful in problems like protein folding, where too many degrees of freedom make complete computation unfeasibly long (Cooper et al., 2010). Another advantage of humans is their ability to watch out for the unexpected. This skill is essential in many realms of science, as evidenced by the discovery of new types of galaxies and astronomical objects in the course of Galaxy Zoo’s operation. Benefits in terms of access to a larger quantity of labor are likely to be greatest for projects involving tasks that are well-structured and require only common human skills, allowing them to draw on a very large pool of potential contributors. This rationale may explain why many existing crowd science projects are located in the bottom-left corner of Figure 4.

Other projects benefit from open participation because it allows them to access rare and specialized skills. Projects such as Argus or eBird (which asks contributors to identify birds in their neighborhoods) can effectively “broadcast” the skill requirement to a large number of individuals in the crowd, enhancing the chances to find suitable contributors. As highlighted by recent work on innovation contests, broadcast search may also identify individuals who already possess pre-existing superior solutions (Jeppesen & Lakhani, 2010; Nielsen, 2012).

Finally, crowd science projects may also benefit from the *diversity* in the knowledge and experience possessed by project participants, especially when problem solutions require the recombination of different pieces of knowledge (Fleming, 2001; Uzzi & Spiro, 2005). Knowledge diversity is likely to be high because project contributors tend to come from different demographic backgrounds, organizations, and even scientific fields (Raddick et al., 2013). Moreover, some projects also benefit from diversity with respect to contributors' geographic location, including efforts such as eBird, Argus, or the Great Sunflower Project, which seek to collect comprehensive data across a large geographic space. In an effort to collect data on the bee population across the United States, for example, The Great Sunflower Project asks participants to grow a special type of flower and to count the number of bee visits during 15-minutes daily observations. The benefits of geographic diversity are likely to accrue primarily to projects requiring observational data where geography itself plays a key role.

4.1.2 Benefits from open disclosure of intermediate inputs

Crowd science relies on the participation of many and often temporary participants. Making intermediate inputs widely accessible enables geographically dispersed individuals to join and to participate via web interfaces. Moreover, codified intermediate inputs such as Polymath blogs or Foldit recipes provide a memory for the project (Frakes & Isoda, 1994; Von Krogh et al., 2003), ensuring that knowledge is not lost when contributors leave and enabling new contributors to quickly catch up and become efficient (Haefliger et al., 2008). Thus, there is an interesting connection between the two dimensions of openness in that the open disclosure of intermediate inputs allows projects to take full advantage of the benefits of open participation (section 4.1.1).

Transparency of the research process and availability of data may also facilitate the verification of results (see Lacetera & Zirulia, 2011). While peer review in traditional science does not typically entail a detailed examination of intermediate inputs, the openness of logs and data in crowd science projects allows observers to follow and verify the research process in considerable detail. By way of example, the relatively large number of "eyes" following the Polymath discussion ensured that mistakes were quickly detected (see Figure 2). Of course, there are limits to this internal verification since contributors who are not trained scientists may not have the necessary background to assess the accuracy of more sophisticated data or methods. As such, their involvement in verification will be limited to tasks similar to the ones they would be able to perform.[20]

A final knowledge-related benefit of open disclosure of intermediate inputs emerges not at the level of a particular project but for the more general progress of science. Scholars of science have argued that open disclosure allows future scientists to build upon prior work in a cumulative fashion (Jones, 2009; Nelson, 2004; Sorenson & Fleming, 2004). While these discussions typically refer to the disclosure

[20] Some projects such as the Open Dinosaur project or Galaxy Zoo institutionalize this process by explicitly asking multiple participants to code the same piece of data and comparing the results.

of final project results, additional benefits may accrue if other scientists can build on the intermediate inputs produced by a project (Dasgupta & David, 1994; Haeussler et al., 2009). Such benefits are particularly large for data, which are typically costly to collect but can often be used to examine multiple research questions. Consider, for example, the Sloan Digital Sky Survey, the Human Genome data, or Census data in the social sciences that have all resulted in many valuable lines of research. Not only data, but also logs and discussion archives can be enormously helpful for future research if they provide insights into successful problem solving techniques. As illustrated in Figure 2, for example, problem solving approaches and intermediate solutions developed in one Polymath discussion have been re-used and adapted in another.

4.2 Motivational benefits

Successful scientific research requires not only knowledge inputs but also scientists who are motivated to actually exert effort towards solving a particular problem. Due to openness in participation, crowd science projects are able to rely on contributors who are willing to exert effort without pay, potentially allowing projects to take advantage of human resources at lower financial cost than would be required in traditional science.[21] In the following, we discuss some of the non-pecuniary payoffs that may matter to crowd science contributors, as well as mechanisms that projects use to address and reinforce these types of motivations. As we will see, some of these mechanisms require considerable infrastructure, suggesting that even though volunteers are unpaid, their help is not necessarily costless.

Scholars have for a long time emphasized the role of intrinsic motives, especially in the context of science and innovation (Hertel et al., 2003; Raasch & Von Hippel, 2012; Ritti, 1968; Sauermann & Cohen, 2010). Intrinsically motivated people engage in an activity because they enjoy the intellectual challenge of a task, because they find it fun, or because it gives them a feeling of accomplishment (Amabile, 1996; Ryan & Deci, 2000). Such intrinsic motives appear to be important also in crowd science projects, as illustrated in the following post of a Galaxy Zoo participant: *"Ok, I'm completely jealous of the person who got that particular image, its amazing. You need to warn people just how addictive this is! Its dangerous! […]. "After doing a couple hundred I was starting to burn out … suddenly there was a kelly-green star in the foreground. Whoa! […] being the first to see these things: who *knows* what you might find? Hooked!"*[22]

Intrinsic motivation may be easy to achieve for tasks that are inherently interesting and challenging. However, even simple and potentially tedious tasks – such as coding large amounts of data

[21] A related argument has been made in the context of user innovation (Raasch & Von Hippel, 2012). While there is no estimate of the cost savings crowd science projects may derive from relying on non-financial motives, there are estimates of the cost savings achieved in OSS development. In particular, one estimate pegs the value of unpaid contributions to the Linux system at over 3 billion dollars (http://linuxcost.blogspot.com/2011/03/cost-of-linux.html; retrieved November 9, 2011).

[22] http://chrislintott.net/2007/07/11/galaxy-zoo-press/

or systematically trying different configurations of molecules – can become intrinsically rewarding if they are embedded in a game-like environment (Prestopnik & Crowston, 2011). As such, an increasing number of crowd science projects employ "gamification" to increase project participation. Figure 1, for example, shows how Foldit uses group and player competitions to keep contributors engaged.

Project participants may also feel good about being part of a particular project community and may enjoy "social" benefits resulting from personal interactions (Hars & Ou, 2002). While some types of projects – especially those involving collaborative problem solving – will naturally provide a locus of social interaction, many projects also actively stimulate interaction. For example, the Great Sunflower project nominates group leaders who operate as facilitators in particular neighborhoods, communities or schools.[23] Other projects promote social interactions by providing dedicated IT infrastructure. This has been the strategy employed by Foldit, where project logs and discussion forums allow participants to team up to exchange strategies and compete in groups, fostering not just enjoyment from gaming, but also a collegial and "social" element.[24] A particularly interesting aspect of intrinsic and social benefits is that they may be non-rival, i.e., the benefits to one individual are not diminished simply because another individual also derives these benefits (Bonaccorsi & Rossi, 2003).

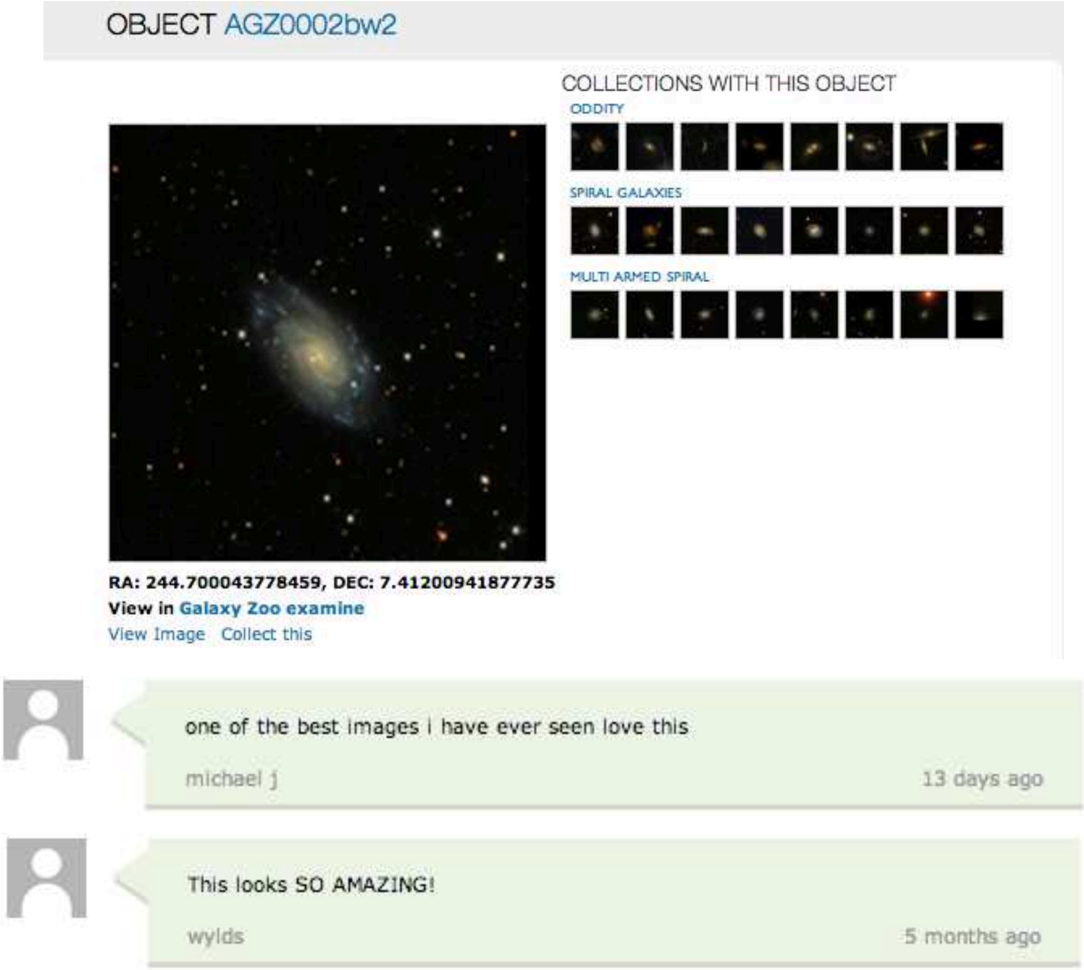
Another important nonpecuniary motive for participation is an interest in the particular subject matter of the project. To wit, when asked about their participation motives, one of the reasons most often mentioned by contributors to Galaxy Zoo was an interest in astronomy and amazement about the vastness of the universe (Raddick et al., 2013). While this motive is intrinsic in the sense that it is not dependent on any external rewards, it is distinct from challenge and play motives in that it is specific to a particular topic, potentially limiting the scope of projects a person will be willing to participate in. Recognizing the importance of individuals' interest in particular topics, some platforms such as Zooniverse enrich the work by providing scientific background information, by regularly informing participants about interesting new findings or by allowing participants to create collections of interesting objects (Figure 5).

While it is not surprising that projects in areas that have long had large numbers of hobbyists – such as astronomy or ornithology – have been able to recruit large numbers of volunteers, projects in less interesting areas, or projects addressing very narrow and specific questions are more likely to face challenges in trying to recruit volunteers. At the same time, reaching out to a large number of potential contributors may allow even projects with less popular topics to identify a sufficient number of individuals with matching interests. As such, while prior work has emphasized that broadcast search can match problems to individuals holding unique knowledge or solutions (Jeppesen & Lakhani, 2010), we suggest that broadcast search can also allow the matching of particular topics to individuals with related

[23] http://www.greatsunflower.org/garden-leader-program; retrieved October 3, 2012.
[24] http://fold.it/portal/blog; retrieved November 11, 2011.

interests. By way of example, bats are probably not popular animals among most people, but the project Bat Detective has been quite successful in finding those individuals who are interested in this species and are willing to listen to sound recordings and to identify different types of bat calls.

Figure 5: Galaxy Zoo object with selected user comments



RA: 244.700043778459, DEC: 7.41200941877735

one of the best images i have ever seen love this

michael j 13 days ago

This looks SO AMAZING!

wylds 5 months ago

Source: http://talk.galaxyzoo.org/objects/AGZ0002bw2

An especially powerful version of interest in a particular problem is evident in the growing number of projects that are devoted to the understanding of particular diseases or to the development of cures (Årdal & Røttingen, 2012). Many of these projects involve patients and their relatives who – as potential "users" – have a very strong personal stake in the success of the project, leading them to make significant time commitments (see Marcus, 2011; Von Hippel, 2006). To these participants, it may be particularly important that projects also quickly disclose intermediate inputs if they believe that disclosure speeds up the progress towards a solution by allowing others to build upon these inputs.

Finally, many crowd science contributors also value the opportunity to contribute to science, especially citizen scientists who lack the formal training to participate in traditional science. In a recent survey of Galaxy Zoo participants, this motive ranked first among 13 alternatives (Raddick et al., 2013). As one Galaxy Zoo contributor commented in a blog discussion: "*Dang … this is addictive. I put on some music and started classifying, and the next thing you know it's hours later. But I'm contributing to* ***science!***"[25] Like some of the other motives discussed above, this motive may be particularly beneficial for crowd science to the extent that individuals are willing to participate in a broad range of different projects and are even willing to help with "boring" and tedious tasks as long as these are perceived to be important for the progress of science.

5 Challenges and potential solutions

Openness with respect to project participation and intermediate inputs can lead to considerable benefits, but the same characteristics may also create certain challenges. We now highlight some of these challenges and draw on the broader organizational literature to point towards organizational and technical tools that may be useful in addressing them.

5.1 Organizational challenges

5.1.1 Matching projects and people

One key feature of crowd science projects is their openness to the contributions of a large number of individuals. However, there is a large and increasing number of projects, and the population of potential contributors is vast. Thus, organizational mechanisms are needed to allow for the efficient matching of projects and potential contributors with respect to both skills and interests. One potential approach makes it easier for individuals to find projects by aggregating and disseminating information on ongoing or planned projects. Websites such as scistarter.com, for example, offer searchable databases of a wide range of projects, allowing individuals to find projects that fit their particular interests and levels of expertise.

We expect that an efficient alternative mechanism may be crowd science platforms that host multiple projects and allow them to utilize a shared pool of potential contributors as well as a shared technical infrastructure. Especially if projects are similar with respect to the field of science, types of tasks, or skill requirements, individuals who contributed to one project are also likely to be suitable contributors to another project on the same platform. Indeed, multi-project crowd science platforms may enjoy considerable network effects by simultaneously attracting more projects looking for contributors and contributors looking for projects to join. Such platforms are common in open source software

[25] http://blogs.discovermagazine.com/badastronomy/2009/04/02/a-million-galaxies-in-a-hundred-hours/#.UU8TPVs4Vxg

development (e.g., sourceforge.com and BerliOS) and are also emerging in the crowd science realm. In particular, the Galaxy Zoo project has evolved into the platform Zooniverse, which currently hosts fifteen projects that cover different areas of science but involve very similar kinds of tasks. When a new project is initiated, Zooniverse routinely contacts its large and growing member base to recruit participants, and many individuals contribute to multiple Zooniverse projects.

5.1.2 Division of labor and integration of contributions

As discussed in section 3.2, projects differ with respect to the complexity and structure of the tasks that are outsourced to the crowd. The more complex and ill-defined the task, the more contributors have to interact and build on each other's contributions, limiting the number of people who can work on a given project at the same time. While our discussion in section 3.2 took the nature of the task as given, project organizers can try to reduce interdependencies and structure tasks such that a larger number of individuals can participate, effectively moving projects towards the left of Figure 4. Open source software development has developed sophisticated modularization techniques to achieve similar goals. The basic idea of modularization is that a large problem is divided into many smaller problems, plus a strategy (the architecture) that specifies how the modules fit together. The goal is to design modules that have minimum interdependencies with one another, allowing for a greater division of labor and parallel work (Baldwin & Clark, 2006; Von Krogh et al., 2003). Modularization has already been used in many crowd science projects. For example, Galaxy Zoo and Old Weather keep the design of the overall project centralized and provide individual contributors with a small piece of the task so that they can work independently and at their own pace. Of course, not all problems can be easily modularized; we suspect, for example, that mathematical problems are less amenable to modularization. However, while the nature of the problem may set limits to modularization in the short term, advances in information technology and project management knowledge are likely to increase project leaders' ability to modularize a given problem over time (see Simon, 1973).

Just as important as distributing tasks is the effective integration of individuals' contributions to find the overall problem solution. In highly modularized data collection and coding tasks such as Old Weather or eBird, individual contributions can easily be integrated into larger data sets. Similarly, in some problem solving tasks such as Foldit, individuals or teams generate stand-alone solutions that can be evaluated using standard criteria and little integration is necessary (see Jeppesen & Lakhani, 2010). The biggest challenge is the integration of contributions in collaborative problem solving tasks such as Polymath, where the contributors seek to develop a single solution in an interactive fashion, e.g., through an ongoing discussion. In such cases, much of the integration is done informally as participants read each other's contributions. However, as the amount of information that needs to be read and processed increases, informal mechanisms may miss important contributions while also imposing large time costs on project participants (Nielsen, 2012). Filtering and sorting mechanisms may lower these costs to some extent, but difficulties in integrating the contributions of a larger number of participants are likely to impose limits upon the optimal size of collaborative problem solving projects such as Polymath.

5.1.3 Project leadership

Most crowd science projects require a significant amount of project leadership. Depending on the nature of the problem, leaders are fundamental in framing the scientific experiment, modularizing the task, securing access to financial and technical resources, or making decisions regarding how to proceed at critical junctures of a project (Mateos-Garcia & Steinmueller, 2008; Wiggins & Crowston, 2011).

Open source software projects such as Linux illustrate that leadership in the form of architecture and kernel design can be performed by a relatively small and tightly knit group of top-notch programmers, while a large number of people at all levels of skills can execute smaller and well-defined modules (Shah, 2006). Emerging crowd science projects such as Galaxy Zoo or Foldit show a similar pattern: These projects are typically led by well-trained scientists who formulate important research questions and design methodologically sound experiments. In collective problem solving projects such as Polymath, leaders are invaluable to wrap up the progress made and keep the project on track. Foldit also illustrates that this kind of leadership is not always exercised in person but can instead be incorporated into technical infrastructure. More specifically, Foldit embeds important "rules of the game" right into the software interface, ensuring that participants perform only operations that project leaders have determined to be consistent with the applicable laws of nature.

While most existing crowd science projects are led by professional scientists, it is conceivable that leadership positions may also be taken by other kinds of individuals. For example, leadership roles might be taken on by designers of collaboration tools, who may have less of an interest in a particular content domain per se, but who have – often through experience – built expertise in crowd science project management. And of course, leaders may emerge from the larger crowd as a project develops. Indeed, the OSS experience suggests that leadership should be thought of as a dynamic concept and can change depending on the particular leadership skills a project requires at a particular point in time (Dahlander & O'Mahony, 2010; O'Mahony & Ferraro, 2007).

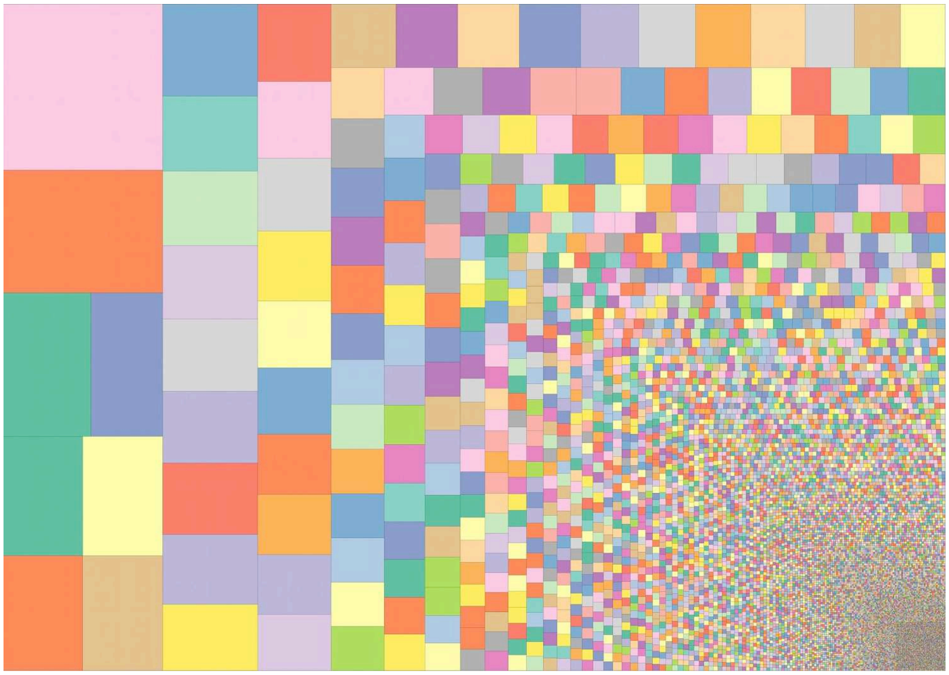
5.2 Motivational challenges

5.2.1 Sustaining contributor involvement

In many crowd science projects, the majority of participants make only small and infrequent contributions, often stopping quickly after joining. To illustrate, Figure 6 shows a square for each of the

16,400 participants in the project Old Weather (as of September, 2012), with the size of the square reflecting the number of classifications made by that individual. We see that a small number of individuals made a very large number of contributions (top left of the figure), while most participants contributed very little (bottom right). Thus, mechanisms to increase the involvement of less frequent contributors may dramatically increase the amount of work a project can get done. At this point, it is not clear what process generates the observed uneven distribution of contributions. One possibility is that most contributors realize shortly after joining that the project is not a good match with respect to their skills or interests (Jovanovic, 1979). However, we suspect that there are also important mechanisms that get people "hooked" over time and through which some of the nonpecuniary motivations discussed earlier become reinforced for some people but not others. Future research on these issues is clearly important.

Figure 6: Distribution of contributions by Old Weather participants



Note: One square for each contributor. Size of square reflects number of contributions per person. Source: http://blog.oldweather.org/2012/09/05/theres-a-green-one-and-a-pink-one-and-a-blue-one-and-a-yellow-one/

5.2.2 Supporting a broader set of motivations

We argued earlier that crowd science projects successfully satisfy certain nonpecuniary motives such as intellectual challenge, interest in a particular area of science, or the desire to contribute to scientific research. These motives are particularly salient to citizen scientists. Yet, projects may also need to find better ways to attract professional scientists or firms because these potential contributors often possess unique capabilities as project organizers, specific kinds of domain specific skills (e.g., top of

Figure 4), or downstream capabilities that are necessary to translate scientific results into useful products or services. While professional scientists and even firms may also be motivated by the nonpecuniary payoffs noted earlier, crowd science projects may face distinct challenges in providing the more "traditional" kinds of payoffs that matter to professional scientists and to firms, mainly due to the high degrees of openness with respect to participation and the disclosure of intermediate inputs as well as research results.

First, the motive to earn peer recognition has always been central in the institution of traditional science (Merton, 1973; Stephan, 2012). To satisfy this motive, crowd science projects need to result in outputs that "count" among professional peers, i.e., publications.[26] While many crowd science projects have resulted in peer-reviewed publications, the assignment of authorship and scientific credit varies dramatically. Some crowd science projects such as Galaxy Zoo, Phylo and Foldit assigned authorship to some of the individual contributors, although such individual credit appears to be reserved primarily for project leaders, potentially limiting the reputational benefits professional scientists can gain when participating as regular contributors. Of course, granting authorship to all contributors - including many citizen scientists who made relatively small contributions and may have little understanding of the scientific problem they helped solve - would dilute the value of authorship and may even violate current authorship guidelines (Haeussler & Sauermann, 2013). Completely eschewing individual credit, some projects name on publications only the group as a whole, such as in the case of "D. H. J. Polymath". The rationale is explained in the Polymath rules:[27] "*If all goes well, a polymath project may end up with one or more publishable results. If so, the paper will be written collaboratively (using the wiki and blog as appropriate), authored under a polymath pseudonym. (The question of assessing how much of a contribution each participant had to a project seems impossible to answer with any degree of objectivity, so the pseudonym approach to authorship is the simplest solution.)*" While the use of group pseudonyms is consistent with the spirit of a collective effort and avoids conflict over authorship attribution, we suspect that it may reduce project participation by professional scientists who need publications to succeed in the traditional institution of science.

Professional scientists' incentives to participate in crowd science may be limited not only because of the way credit is assigned. Perhaps more importantly, the open disclosure of intermediate inputs allows competing projects to build on such knowledge more intensively and faster than would be the case in traditional science, potentially hurting a crowd science projects' ability to win the race for priority in discovery. While competition among crowd science projects pursuing the same problem is currently rare,

[26] Similar career related motives have also been highlighted in prior work on the motivations of open source software contributors (Lerner & Tirole, 2005).
[27] http://polymathprojects.org/general-polymath-rules/

such competition may increase as crowd science grows and project platforms proliferate. In some domains, crowd science projects may also compete directly with projects in traditional science. Thus, while the open disclosure of intermediate inputs has several benefits (section 4.1.2), it may also create a competitive disadvantage and weaken professional scientists' incentives to get involved.

A second motive that is currently difficult for crowd science to satisfy is money - direct financial rewards for project participation are largely absent. This low importance of financial rewards follows quite directly from the high degrees of openness with respect to both participation and knowledge disclosure. With respect to the former, the large number of contributors makes financial payments logistically challenging and very costly. More importantly, since crowd science projects disclose openly both the final project results and intermediate inputs, their ability to appropriate significant financial returns will be limited (see Cohen et al., 2000), leaving little money to be distributed to contributors. Given the success current projects have had without providing pay to participants, one may argue that financial incentives are unnecessary. However, we expect that finding ways to address financial motives may be important to attract individuals with particular rare skills or pre-existing solutions that are highly valued in the traditional labor market or that promise high returns when used in innovation contests that promise large payouts (quadrant 1 in Figure 3). It would be conceivable that projects use grant money to compensate not just project leaders but also key contributors with highly valued skills or knowledge. Of course, any such mechanism would face challenges with respect to maintaining perceptions of fairness, and avoiding that the nonpecuniary motivations of regular contributors are undermined.

The expectation of at least some level of financial returns may also be important to firms. Firm involvement may be beneficial for crowd science projects because some firms command significant human resources that would be valuable to projects requiring domain-specific skills.[28] Firms may also be essential as partners who possess other assets that are essential for the ultimate success of a crowd science project, e.g., the complementary capabilities required to develop and market a new drug compound that has been identified through a crowd science effort (Årdal & Røttingen, 2012). While firm involvement in crowd science is rare – presumably because of the absence of sufficient financial incentives – the open source software experience has shown that firms may find profitable ways to get involved even in projects that openly disclose intermediate inputs and final results by developing proprietary value-added or by selling products that draw upon open source outputs (Hars & Ou, 2002; Lakhani & Wolf, 2006; Lerner & Tirole, 2005). Firms may find similar ways to benefit from crowd science, potentially in the context of broader open innovation strategies (Chesbrough, 2003; Dahlander & Gann, 2010). Indeed, an interesting example of a company-run crowd science project is Argus, which asks captains to collect data on ocean

[28] There is an interesting parallel in the open source software space: many OSS developers contribute as paid employees of firms such as IBM or Intel (Hars & Ou, 2002; Lakhani & Wolf, 2006).

depth and incorporates these data in freely available maps that can be used for navigation. At the same time, the sponsoring firm – Survice Engineering – incorporates the data in its more sophisticated commercial maritime products.

5.2.3 Reconciling conflicting motivations

If crowd science projects seek to attract participants with different types of motives, conflict may arise (see Harhoff & Mayrhofer, 2010). However, the open source software experience suggests that different motivations can co-exist within a project, and that potential incentive conflicts can be mitigated using contractual mechanisms (Lerner & Tirole, 2005; McGowan, 2001). The key insight is that it is useful to distinguish different rights associated with the collective production of knowledge including, for example, the right to be regarded as the author, the right of using a product that was collectively produced, or the right to make derived works. Unbundling these rights can help in attracting diverse groups of individuals because at least some of these rights are not strictly rival and can be enjoyed at the same time (Bonaccorsi & Rossi, 2003).

To illustrate, let us consider the example of Solar Stormwatch. This project asks people to watch videos of solar storms and to tag characteristics such as the inception point or the maximum reach of the storm. The small group of lead scientists on this project includes professional scientists working at a government lab as well as a PhD student at Imperial College, London.[29] Let us assume that these scientists are primarily motivated by the desire to write scientific papers based on the data resulting from the project. Suppose now that a company producing solar panels for satellites wants to use these data to enable its equipment to detect the inception of a solar storm. The company may be willing to participate in the project to speed up the completion and release of the dataset, e.g., by paying an employee to work on the project. However, it will do so only if it is ensured access to the resulting data and if it can incorporate the data into its proprietary computer algorithms. There is little conflict between the company's plans and scientists' desire to publish the data or papers based on the data. Now consider a third party, namely another team of astronomers who need data on solar activity for their own research. These researchers may be willing to help with the project if they are ensured open and timely access to the data. However, if these researchers are working on a similar problem as the Solar Stormwatch lead scientists, the two teams are directly competing, potentially reducing their incentives to invest effort in the project. In contrast, both teams of scientists should be willing to participate if they expect to use the data to pursue non-competing research questions (see Von Hippel & Von Krogh, 2003).[30] Finally, consider a fourth set of contributors – citizen scientists who simply enjoy watching videos of solar storms and

[29] http://www.solarstormwatch.com/mission_briefing; retrieved 12 February 2012.

[30] This discussion suggests that crowd science projects may be more viable in "general" research areas that allow the re-use of data for several non-competing research streams.

learning more about this exciting phenomenon. In principle, this latter group of contributors should not care who gets credit for scientific results from the project, and they should also not be opposed to a company creating useful products based on the resulting knowledge.

Taking inspiration from existing open source software license arrangements, contractual mechanisms can be envisioned to incentivize all parties in our example to participate while mitigating potential goal conflicts. For example, the founding team could reserve the right to use the data for particular pre-defined research questions or for a limited amount of time, ensuring that the lead scientists have incentives to invest the time and resources required to run the project. At the same time, the lead scientists would commit to disclose the data openly for other uses – providing incentives for the second team of professional scientists.[31] The license could also specify that algorithms derived from the data may be incorporated in commercial products, incentivizing the firm to participate in the project, without reducing the incentives for either group of scientists. Citizen scientists primarily need the permission to use the project infrastructure – which the organizers should willingly provide in return for free help.

While the open source experience suggests that well-defined and modular rights can be central to making community-based production a stable phenomenon, such contractual arrangements may also face various challenges in the context of crowd science. Among others, it is not clear how well license contracts can be enforced, although some crowd science platforms such as Zooniverse already require contributors to log in and accept certain terms and conditions. Another important concern is that such arrangements undermine some of the openness that distinguish crowd science projects from other knowledge production regimes and that may explain their initial success. Overall, a delicate balance seems needed as projects seek ways to accommodate the desires of project participants to derive certain kinds of rival rewards, while also reaping the knowledge-related benefits of openness. Given the ad hoc nature of some of the mechanisms currently employed, future systematic research on the costs and benefits of contractual arrangements in the specific context of crowd science would clearly be important.

6 Crowd science: A research agenda

To the extent possible, our discussion in the foregoing sections was based on qualitative evidence from a limited number of existing crowd science projects, as well as the small body of empirical work that has started to investigate crowd science more systematically. In addition, we built on related streams of literature in organizational theory as well as on open source software development. Many of the

[31] Indeed, such an agreement is spelled out in the frequently asked questions section of the Open Dinosaur Project: "***May I use the data for my own research?*** *Yes, you may – although we ask that you hold off on publication of any results until we have had a chance to complete the initial publication of our own study. The reason for this is that we feel it is important for our volunteers and project leaders to get their chance at the rewards of this project (which have very real implications for tenure, promotion, graduate school admission, etc.) without being scooped. Once we have published the initial results, all of the data incorporated into the study are completely fair game.*" Source: http://opendino.wordpress.com/faqs/; retrieved April 8, 2013.

conjectures developed in our discussion provide fertile ground for future qualitative and quantitative research. For example, research is needed on the degree to which multi-project platforms can improve the matching between projects and contributors, or allow for an efficient use of technical infrastructure. Similarly, future work is needed to gain insights into the relative importance of various types of motivations and into the degree to which crowd science projects experience conflicts among contributors. In the following, we point towards some broader questions for future research that were less salient in our discussion but that may be just as important.

First, some observers have expressed concerns regarding the scalability of the crowd science approach. After all, if 700,000 people participate on the Zooniverse platform, how many people are left to participate in other crowd science projects? Galaxy Zoo's Chris Lintott appears relaxed, stating: "*We have used just a tiny fraction of the human attention span that goes into an episode of Jerry Springer.*" (quoted in Cook, 2011). At the minimum, it will be important to understand how projects can expand beyond the relatively small body of "early adopters" to involve broader segments of the populations of professional scientists and potential citizen scientists (including Jerry Springer fans). Crowston and Fagnot (2008) begin to develop a dynamic model of virtual collaboration that may be useful in thinking about this question.

Second, while much of the current discussion focuses on how crowd science projects form and operate, very little is known regarding the quantity and quality of research outputs. One particularly salient concern is that projects that are initiated by non-professional scientists may not follow the scientific method, calling in question the quality of research output. Some citizen science projects led by patients, for example, do not use the experimental designs typical of traditional studies in the medical sciences, making it difficult to interpret the results (Marcus, 2011). To ensure that crowd science meets the rigorous standards of science, it seems important that trained scientists are involved in the design of experiments. To some extent, however, rigor and standardized scientific processes may also be embedded in the software and platform tools that support a crowd science project. Similarly, it may be possible for crowd science platforms to provide "scientific consultants" who advise (and potentially certify) citizen science projects. Finally, to the extent that crowd science results are published in traditional journals, the traditional layer of quality control in the form of peer review still applies. However, results are increasingly disclosed through non-traditional channels such as blogs and project websites. The question whether and how such disclosures should be verified and certified is an important area for future scholarly work and policy discussions.

A related question concerns the efficiency of the crowd science approach. While it is impressive that the Zooniverse platform has generated dozens of peer reviewed publications, this output does not reflect the work of a typical academic research lab. Rather, it reflects hundreds of thousands of hours of labor supplied by project leaders as well as citizen scientists (see a related discussion in Bikard & Murray, 2011). Empirical research is needed to measure crowd science labor inputs, possibly giving different weights to different types of skills (see Figure 4). It is likely that most crowd science projects are less efficient than traditional projects in terms of output relative to input; however, that issue may be less of a concern given that most of the labor inputs are provided voluntarily and for "free" by contributors who appear to derive significant non-pecuniary benefits from doing so. Moreover, some large-scale projects would simply not be possible in a traditional lab. Nevertheless, understanding potential avenues to increase efficiency will be important for crowd science's long-term success. By way of example, the efficiency of distributed data coding projects such as Galaxy Zoo may be increased by tracking individuals' performance over time and limiting the replication of work done by contributors who have shown reliable performance in the past (see Simpson et al., 2012).

As shown in Figure 4, most existing crowd science projects involve well-structured tasks of low complexity, many of which involve the collection or coding of data. While such projects have resulted in important scientific insights, a key question is whether and how the crowd science approach can be leveraged for tasks that are currently more complex and ill-defined. Examples such as Foldit and Polymath show that the crowd can successfully work on those types of tasks, yet we have also discussed some of the organizational challenges that are likely to arise. As such, future research on modularization and other mechanisms that allow involvement of the crowd at a larger scale would be of particularly great value.

Our study is inductive in that we started with observations of existing crowd science projects, provided a discussion of their key features, and sought to understand potential benefits and challenges by drawing on existing work in other research areas as a conceptual guide. As our understanding of crowd science advances, future research may develop a theoretical framework to examine under which conditions or for which kinds of projects crowd science is a superior organizational mode compared to alternatives. This approach has recently been employed to study optimal governance choice for problem solving in firms (Afuah & Tucci, 2012; Felin & Zenger, 2012). That research has provided important insights by focusing primarily on characteristics of the problem (such as complexity or distance from existing knowledge), yet our discussion suggests that future work should consider not just knowledge-related aspects of scientific research but also motivational aspects. Moreover, it seems insufficient to consider only the strength of abstractly defined incentives since the richness and diversity of pecuniary and especially non-pecuniary motivations makes crowd science and other forms of community-based knowledge production so interesting and potentially powerful (see also Lerner & Tirole, 2005; Von Hippel, 2006; Von Krogh et al., forthcoming).

Finally, while our discussion of future research has focused on crowd science as the object of study, crowd science may also serve as an ideal setting to study a range of issues central to our understanding of science and knowledge production more generally. For example, the team size in traditional science has been increasing in most fields (Wuchty et al., 2007), raising challenges associated with the effective division of labor and the coordination of project participants (see Cummings & Kiesler, 2007). As such, research on the effective organization of crowd science projects may also inform efforts to improve the efficiency of (traditional) team science. Similarly, crowd science projects may provide unique insights into the process of knowledge creation. For example, detailed discussion logs may allow scholars to study cognitive aspects of problem solving and the interactions among individuals in scientific teams (Singh & Fleming, 2010), or to compare the characteristics of both successful and unsuccessful problem solving attempts. Such micro-level insights are extremely difficult to gain in the context of traditional science, where disclosure is limited primarily to the publication of (successful) research results, and where the path to success remains largely hidden from the eyes of social scientists.

7 Conclusion and policy implications

At the beginning of this paper, we introduced the reader to crowd science by describing three prominent examples of crowd science projects. We then developed a conceptual framework to characterize crowd science and distinguish it from other regimes of knowledge production. In doing so, we highlighted two features: openness with respect to project participation and openness with respect to intermediate inputs. We proceeded by discussing potential benefits and challenges resulting from these characteristics and conjectured how some of the challenges may be addressed. We then outlined an agenda for future research on crowd science itself, while also highlighting potential benefits of using crowd science as empirical setting to study knowledge production processes more generally. We now conclude by considering potential implications for policy makers and funding agencies.

While much research remains to be done on specific aspects of crowd science, the success of existing projects suggests that crowd science can make significant contributions to science and deserves the attention of funding agencies and policy makers. Indeed, crowd science may be particularly appealing to funding agencies for several reasons. First, by complementing the time of lead researchers and costly physical resources with (unpaid) contributions from the larger crowd, crowd science projects may yield higher returns to a given monetary investment than projects in traditional science. In addition, by disclosing intermediate inputs, crowd science projects may provide greater "spillovers" to other projects and generate greater benefits for the general progress of science than projects that only publish final results. As noted earlier, funding agencies are keenly aware of such benefits and are increasingly mandating disclosure of intermediate results in traditional science, although the resulting disclosure is likely less comprehensive than in crowd science projects. Finally, many crowd science projects involve citizen scientists, potentially increasing the public's understanding of scientific activity and of the value of publicly funded research.

To the extent that funding agencies are interested in supporting crowd science, investments in crowd science infrastructure may be particularly useful. Such infrastructure may include crowd science platforms that host multiple projects (e.g., Zooniverse), thus lowering the cost of starting new projects. In a more general sense, such "infrastructure" may also entail organizational and management knowledge resulting from social sciences research into the effective organization of crowd science projects. Finally, funding support may be needed to preserve intermediate inputs that are disclosed by crowd science projects but are not systematically archived by traditional journals or libraries. This potentially valuable resource is at risk to be lost when projects are completed and participants re-dedicate their time and project infrastructure to new projects.[32] Funding agencies as well as policy makers may also play an important role in discussing and coordinating the development of standardized licenses or other contractual mechanisms that may allow projects to govern the collaboration among heterogeneous sets of project participants. As discussed in section 5.2.3, the open source software experience suggests that such tools can foster the development of community-based production and may be particularly useful in reconciling potentially conflicting motives of different types of participants.[33] Finally, funding agencies, policy makers, and scholarly organizations should engage in discussions regarding how the quality of research can be assured in projects that do not involve professionally trained scientists and that use the Internet to disclose research results and data without the use of traditional journals and the associated process of peer review.

[32] http://blogs.loc.gov/digitalpreservation/2012/07/preserving-online-science-reflections/

[33] See for example the repository of the Open Source initiative, http://www.opensource.org/osd.html

REFERENCES

Afuah, A., & Tucci, C. L. 2012. Crowdsourcing as a solution to distant search. ***Academy of Management Review***, 37(3): 355-375.

Amabile, T. 1996. ***Creativity in Context***. Boulder, Colo.: Westview Press.

Årdal, C., & Røttingen, J. A. 2012. Open source drug discovery in practice: A case study. ***PLOS Neglected Tropical Diseases***, 6(9): e1827.

Baldwin, C. Y., & Clark, K. B. 2006. The architecture of participation: Does code architecture mitigate free riding in the open source development model? ***Management Science***, 52(7): 1116.

Bikard, M., & Murray, F. 2011. Is collaboration creative or costly? Exploring tradeoffs in the organization of knowledge work., ***Working Paper***.

Bonaccorsi, A., & Rossi, C. 2003. Why Open Source software can succeed. ***Research Policy***, 32(7): 1243-1258.

Boudreau, K. J., Lacetera, N., & Lakhani, K. R. 2011. Incentives and problem uncertainty in innovation contests: An empirical analysis. ***Management Science***, 57(5): 843-863.

Brokaw, L. 2011. Could "citizen science" be better than academy science? ***MIT Sloan Management Review Blog***.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C., Lintott, C., Keel, W. C., Parejko, J., & Nichol, R. C. 2009. Galaxy Zoo Green Peas: discovery of a class of compact extremely star−forming galaxies. ***Monthly Notices of the Royal Astronomical Society***, 399(3): 1191-1205.

Chesbrough, H. W. 2003. ***Open Innovation: The New Imperative for Creating and Profiting from Technology***. Boston, Mass.: Harvard Business School Press.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. 2000. Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not), ***NBER Working Paper #7552***.

Cook, G. 2011. How crowdsourcing is changing science, ***The Boston Globe***.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., & Popovic, Z. 2010. Predicting protein structures with a multiplayer online game. ***Nature***, 466(7307): 756-760.

Crowston, K., & Fagnot, I. 2008. The motivational arc of massive virtual collaboration, ***Working Paper***.

Cummings, J. N., & Kiesler, S. 2007. Coordination costs and project outcomes in multi-university collaborations. ***Research Policy***, 36(10): 1620-1634.

Dahlander, L., & Gann, D. M. 2010. How open is innovation? ***Research Policy***, 39(6): 699-709.

Dahlander, L., & O'Mahony, S. 2010. Progressing to the center: Coordinating project work. ***Organization Science***, 22(4): 961-979.

Dasgupta, P., & David, P. A. 1994. Toward a new economics of science. ***Research Policy***, 23(5): 487-521.

David, P. 2008. The historical origins of "open science". ***Capitalism and Society***, 3(2).

Erickson, L., Petrick, I., & Trauth, E. 2012. Hanging with the right crowd: Matching crowdsourcing need to crowd characteristics. ***Proceedings of the 8th Americas Conference on Information Systems***.

Felin, T., & Zenger, T. R. 2012. Open innovation, problem-solving and the theory of the (innovative) firm, ***Working Paper***.

Fleming, L. 2001. Recombinant uncertainty in technological search. ***Management Science***, 47(1): 117-132.

Frakes, W. B., & Isoda, S. 1994. Success factors of systematic reuse. ***Software, IEEE***, 11(5): 14-19.

Haefliger, S., Von Krogh, G., & Spaeth, S. 2008. Code reuse in open source software. ***Management Science***, 54(1): 180-193.

Haeussler, C., Jiang, L., Thursby, J., & Thursby, M. 2009. Specific and general information sharing among academic scientists, ***NBER Working Paper #15315***.

Haeussler, C., & Sauermann, H. 2013. Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. ***Research Policy***, 42(3): 688-703.

Harhoff, D., & Mayrhofer, P. 2010. Managing user communities and hybrid innovation processes: Concepts and design implications. ***Organizational Dynamics***, 39(2): 137-144.

Hars, A., & Ou, S. 2002. Working for free? Motivations for participating in Open-Source projects. ***International Journal of Electronic Commerce***, 6(3): 25-39.

Hertel, G., Niedner, S., & Herrmann, S. 2003. Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. ***Research Policy***, 32(7): 1159-1177.

Jeppesen, L. B., & Lakhani, K. 2010. Marginality and problem-solving effectiveness in broadcast search. ***Organization Science***, 21(5): 1016-1033.

Jones, B. 2009. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? ***Review of Economic Studies***, 76(1): 283-317.

Jovanovic, B. 1979. Job matching and the theory of turnover. ***The Journal of Political Economy***, 87(5): 972-990.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., Baker, D., & Players, F. 2011a. Algorithm discovery by protein folding game players. ***Proceedings of the National Academy of Sciences***, 108(47): 18949-18953.

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., & Popović, Z. 2011b. Crystal structure of a monomeric retroviral protease solved by protein folding game players. ***Nature Structural & Molecular Biology***, 18(10): 1175-1177.

Kuhn, T. S. 1962. ***The Structure of Scientific Revolutions***: University of Chicago Press.

Lacetera, N., & Zirulia, L. 2011. The economics of scientific misconduct. ***Journal of Law, Economics, and Organization***, 27: 568-603.

Lakhani, K., & Wolf, R. 2006. Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. In J. Feller, B. Fintzgerald, S. Hissam, & K. Lakhani (Eds.), ***Perspectives on Free and Open Source Software***: MIT Press.

Lerner, J., & Tirole, J. 2005. The economics of technology sharing: Open source and beyond. ***Journal of Economic Perspectives***, 19(2): 99.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., & Raddick, M. J. 2010. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. ***Monthly Notices of the Royal Astronomical Society***.

Malone, T. W., & Klein, M. 2007. Harnessing collective intelligence to address global climate change. ***Innovations: Technology, Governance, Globalization***, 2(3): 15-26.

Marcus, A. D. 2011. Citizen scientists, ***Wall Street Journal***.

Mateos-Garcia, J., & Steinmueller, E. 2008. Open, but how much? Growth, conflict, and institutional evolution in open-source communities., ***Community, Economic Creativity, and Organization***: 254-281: Oxford University Press.

McGowan, D. 2001. Legal implications of open-source software. ***University of Illinois Law Review***: 241.

Merton, R. K. 1973. ***The Sociology of Science: Theoretical and Empirical Investigations***. Chicago: University of Chicago Press.

Murray, F., & O'Mahony, S. 2007. Exploring the foundations of cumulative innovation: Implications for Organization Science. ***Organization Science***, 18(6): 1006-1021.

Nelson, R. 2004. The market economy, and the scientific commons. ***Research Policy***, 33(3): 455-471.

Nickerson, J. A., & Zenger, T. R. 2004. A knowledge-based theory of the firm: The problem-solving perspective. ***Organization Science***, 15(6): 617-632.

Nielsen, M. 2012. ***Reinventing Discovery: The New Era of Networked Science***: Princeton University Press.

O'Mahony, S., & Ferraro, F. 2007. The emergence of governance in an open source community. ***The Academy of Management Journal***, 50(5): 1079-1106.

Polymath, D. H. J. 2012a. Deterministic method to find primes. ***Mathematics of Computation***, 81(278): 1233-1246.

Polymath, D. H. J. 2012b. A new proof of the density Hales-Jewett theorem. ***Annals of Mathematics***, 175: 1283-1327.

Prestopnik, N. R., & Crowston, K. 2011. Gaming for (Citizen) Science: Exploring Motivation and Data Quality in the Context of Crowdsourced Science through the Design and Evaluation of a Social-Computational System: 28-33: IEEE.

Raasch, C., & Von Hippel, E. 2012. Amplifying user and producer innovation: The power of participation motives, ***Working Paper***.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C., Cardamone, C., Murray, P., Schawinski, K., Szalay, A., & Vandenberg, J. 2013. Galaxy Zoo: Motivations of Citizen Scientists, ***Working Paper***.

Ritti, R. 1968. Work goals of scientists and engineers. ***Industrial Relations***, 8: 118-131.

Ryan, R. M., & Deci, E. L. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. ***Contemporary Educational Psychology***, 25(1): 54-67.

Sauermann, H., & Cohen, W. 2010. What makes them tick? Employee motives and firm innovation. ***Management Science***, 56(12): 2134-2153.

Sauermann, H., & Stephan, P. 2013. Conflicting logics? A multidimensional view of industrial and academic science. ***Organization Science***.

Shah, S. K. 2006. Motivation, governance, and the viability of hybrid forms in open source software development. ***Management Science***, 52(7): 1000-1014.

Simon, H. A. 1962. The architecture of complexity. ***Proceedings of the American Philosophical Society***: 467-482.

Simon, H. A. 1973. The structure of ill structured problems. ***Artificial Intelligence***, 4(3): 181-201.

Simpson, E., Roberts, S., Psorakis, I., & Smith, A. 2012. Dynamic Bayesian combination of multiple imperfect classifiers, ***arXiv Working Paper***.

Singh, J., & Fleming, L. 2010. Lone inventors as sources of breakthroughs: Myth or reality? ***Management Science***, 56(1): 41-56.

Sorenson, O., & Fleming, L. 2004. Science and the diffusion of knowledge. ***Research Policy***, 33(10): 1615-1634.

Stephan, P. 2012. ***How Economics Shapes Science***: Harvard University Press.

Stephan, P. E. 1996. The economics of science. ***Journal of Economic Literature***, 34(3): 1199-1235.

Uzzi, B., & Spiro, J. 2005. Collaboration and creativity: The small world problem. ***American Journal of Sociology***, 111(2): 447-504.

Von Hippel, E. 2006. ***Democratizing Innovation***: MIT Press.

Von Hippel, E., & Von Krogh, G. 2003. Open source software and the" private-collective" innovation model: Issues for organization science. ***Organization Science***: 209-223.

Von Krogh, G., Haefliger, S., Spaeth, S., & Wallin, M. W. forthcoming. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. ***MIS Quarterly***.

Von Krogh, G., Spaeth, S., & Lakhani, K. R. 2003. Community, joining, and specialization in open source software innovation: A case study. ***Research Policy***, 32(7): 1217-1241.

Walsh, J. P., Cho, C., & Cohen, W., M. . 2005. View from the bench: Patents and material transfers. ***Science***, 309(5743): 2002-2003.

Weisberg, R. W. 2006. ***Creativity: Understanding Innovation in Problem Solving, Science, Invention, and the Arts***: Wiley.

Wiggins, A., & Crowston, K. 2011. ***From conservation to crowdsourcing: A typology of citizen science***. Paper presented at the 44th Hawaii International Conference on Systems Sciences (HICSS).

Wuchty, S., Jones, B., & Uzzi, B. 2007. The increasing dominance of teams in the production of knowledge. ***Science***, 316(5827): 1036-1039.

Young, J. 2010. Crowd science reaches new heights. ***The Chronicle of Higher Education***, 28.

Zuckerman, H. 1988. The sociology of science. In N. J. Smelser (Ed.), ***The Handbook of Sociology***: 511-574: Sage.

Table 1: Examples of Crowd Science Projects

NAME	URL	FIELD	PRIMARY TASK
Ancient Lives	http://ancientlives.org	Archeology	Transcribe
Argus	http://argus.survice.com/	Oceanology	Measure & input
Bat Detective	http://www.batdetective.org/	Zoology	Listen & classify
Cyclone Center	http://www.cyclonecenter.org/	Climatology	Classify
Discovery Life	http://www.discoverlife.org/pa/ph/	Biology	Input
eBird	www.ebird.org	Zoology	Observe & input
Field Expedition-Mongolia	http://exploration.nationalgeographic.com/mongolia	Archeology	Identify & flag
Eterna	http://eterna.cmu.edu/	Biochemistry	Game
Foldit	www.fold.it	Biochemistry	Game
Galaxy Zoo	www.galaxyzoo.org	Astronomy	Classify & flag
Great Sunflower Project	www.greatsunflower.org	Biology	Plant, observe & input
Ice Hunters	http://www.icehunters.org	Astronomy	Identify & flag
Moon Zoo	www.moonzoo.org	Astronomy	Identify & flag
Old Weather	http://www.oldweather.org	Climatology	Transcribe
Open Source Drug Discovery/C2D project	http://c2d.osdd.net/	Drug discovery	Annotate
Open Dinosaur Project	http://opendino.wordpress.com/about/	Paleontology	Input
Patientslikeme	http://www.patientslikeme.com/	Medicine	Input
Pigeon Watch	http://www.birds.cornell.edu/pigeonwatch	Ornithology	Input
Phylo	http://phylo.cs.mcgill.ca	Genetics/ bioinformatics	Game
Planet Hunters	http://www.planethunters.org	Astronomy	Classify & flag
Polymath	www.polymathprojects.org	Mathematics	Problem solving
Seafloor Explorer	http://www.seafloorexplorer.org/	Marine biology	Identify & flag
Seti@home	https://www.zooniverse.org/lab/setilive	Space exploration	Identify & flag
SetiQuest	http://setiquest.org/	Astronomy	Identify & flag
SOHO Comet Hunting	http://scistarter.com/project/529-SOHO%20Comet%20Hunting	Astronomy	Identify & flag
Solar Stormwatch	http://www.solarstormwatch.com	Astronomy	Identify & flag
Space NEEMO	https://www.zooniverse.org/lab/neemo	Marine biology	Identify & flag
Stardust@home	http://stardustathome.ssl.berkeley.edu/	Astronomy	Identify & flag
Synaptic Leap Schistosomiasis project	http://www.thesynapticleap.org/schist/projects	Pharmacology	Experiment
Whale Song	http://whale.fm	Zoology	Listen & match
What's the score	http://www.whats-the-score.org/	Music	Transcribe