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 knowledge - and distinguish crowd science from other knowledge production regimes. Third, we explore potential benefits that crowd science offers over alternative organizational modes, and potential challenges it is likely to face. This discussion also considers for what kinds of scientific problems 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; open science

November 14, 2012

We thank Pierre Azoulay, Rhiannon Crain, Carolin Haeussler, Firas Khatib, Karim Lakhani, Danielle Li, Chris Lintott, Ali Mohammadi, Alex Oettl, Cristina Rossi-Lamastra, Arfon Smith, Ed Steinmueller and Jose Urbina for comments and suggestions. We are also grateful for many stimulating discussions at the 10th International Open and User Innovation Workshop and the 2012 NBER Workshop on Scholarly Communication, Open Science, and its Impacts. 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, consider Polymath, 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; Raddick et al., 2009; 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 knowledge 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 the potential benefits and challenges of crowd science.[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*.[2] 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 project participants. A deeper understanding of these various benefits and challenges may allow us to assess the general

[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.

[2] Moreover, crowd science projects may help promote an interest in science in the broader population. As such, several crowd science projects are funded in the U.S. under programs for advancing learning such as the Advancing Informal STEM Learning program (AISL) http://www.nsf.gov/news/news_summ.jsp?cntn_id=124991&org=NSF, retrieved November 7, 2012.

prospects of crowd science, but also to 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 distinguishing them from projects in "traditional science" and other emerging organizational paradigms such as crowd sourcing and innovation contests. Third, we draw on the broader organizational literature to discuss potential benefits and challenges crowd science projects are likely to face, and conjecture how these challenges may be addressed. This discussion also considers for what kinds of scientific problems the crowd science approach may be particularly suitable. Finally, we outline an agenda for future research and discuss potential implications for funding agencies and policy makers.

2 Examples of crowd science projects

2.1 Foldit[3]

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 include X-ray crystallography, nuclear magnetic resonance spectroscopy, and electron microscopy. Unfortunately, these methods are extremely expensive, costing up to 100,000 dollars for a single protein and there are 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 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

[3] 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.

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 combines 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.

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.

In the months to come 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 had set up teams with names such as "Another hour, another point" or "Void Crushers" and posted strategies for protein folding. Chat and forum were always active. Some players had begun to elaborate their own "recipes", encoded strategies that could be compared to those created in the lab. And some of 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, 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 the group of Chris Lintott 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

[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.

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 some online tutorials, 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 millions galaxies had been coded and multiple classifications by different volunteers were used to reduce the incidence of incorrect coding. For an individual scientist, this task 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 Knowedge currently shows eight published papers, and a search for "Green pea galaxies" gives six.

The coded Galaxy Zoo data were made 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 coding 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 under a common platform "The

[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.

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 scientists to save precious observation time at large telescopes. Lintott 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 one of the best living 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". In the next hours, several readers commented on his post. 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. 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

[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.

significant progress was coming form various sources. On March 10, after six weeks of discussion, Gowers announced that 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 sent for publication to the Annals of Mathematics under the pseudonym of "D.H.J. Polymath".

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. These 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, 2012). 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 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.[9]

It is interesting to note that over time, Polymath projects have developed certain practices that facilitate the discussions. In particular, a common problem 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, discouraging new people from joining and 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, provide answers or rate others' answers in a commented blog-style discussion. In some cases, the discussion develops in ways similar to those of a small collective Polymath project and answers have often been cited in scholarly articles.

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

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 kinds of tasks are most common in a particular type of project, 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 which these projects differ from other emerging approaches to producing knowledge such as crowd sourcing or innovation contests. In the following section, we identify two key features that appear to distinguish crowd science from other regimes of knowledge production: openness in project participation and openness with respect to the disclosure of intermediate project results and data. We do not claim that all projects share these features to the same degree. However, these dimensions tend to distinguish crowd science from other organizational forms, while also potentially having important implications for opportunities and challenges crowd science projects may face. In the subsequent section, 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 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. Individuals who are interested in a project are free to join, even if they can only make small time commitments. Moreover, many types of crowd science tasks do not require scientific training and attract contributors from the general population – explaining the frequent use of the term

"citizen science". 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 openly 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.

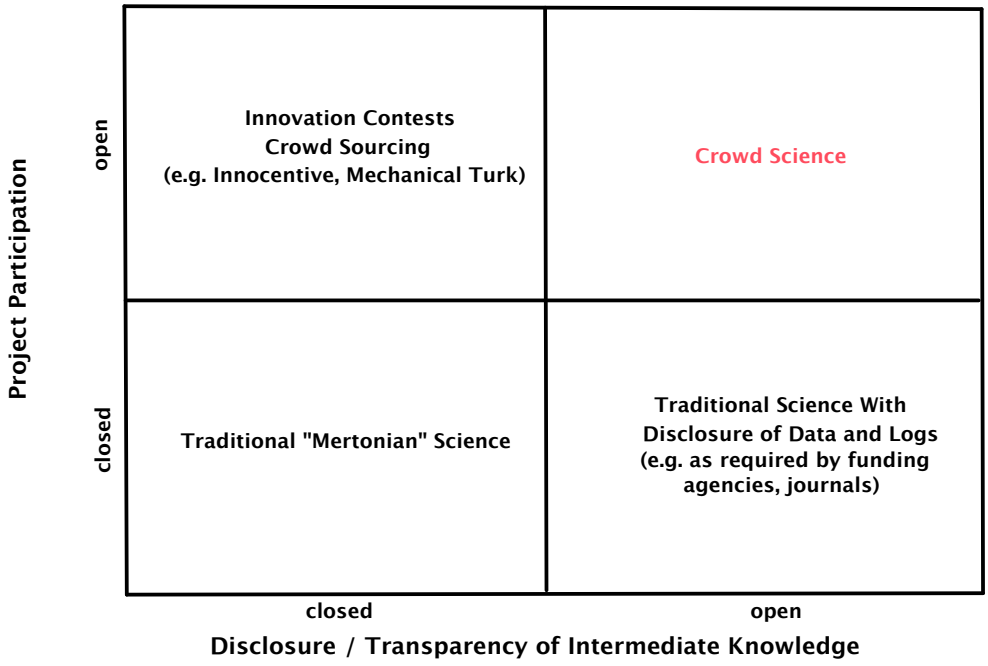
A second feature that tends to be common in crowd science projects is that they openly disclose at least some of the "intermediate knowledge" that arises over the course of the project. We use the term intermediate knowledge in a very broad sense. Such intermediate knowledge may include raw data such as images, videos, and audio-tracks that are to be coded or otherwise utilized by project participants. The Whale Song Project, for example, has uploaded a database of audio recordings of whale songs, obtained by combining various independent sources of recordings. Many projects also provide detailed insights into the actual process of problem solving. For example, Polymath blogs preserve the ongoing discussions among project participants, keeping track of both successful and unsuccessful problem solving attempts. Similarly, other platforms such as Foldit and Galaxy Zoo include discussion boards that allow contributors to exchange ideas, to share unexpected findings (such as "Annie's vorveerp"), or to seek advise from fellow contributors. Finally, projects may disclose data that resulted from contributors' work for re-use in subsequent projects. For example, the classifications of galaxies contributed by Galaxy Zoo volunteers were made available publicly in the summer of 2010 (Lintott et al., 2010).

The high degree of openness with respect to intermediate knowledge is critical for the functioning of crowd science projects. Like open source software (OSS) development, crowd science relies on the participation of many and often temporary participants. Given frequent turnover, the codified knowledge recorded in the system as a whole provides the memory to ensure continuity in a project (Frakes & Isoda, 1994; Von Krogh et al., 2003), enabling new entrants to quickly catch up and become efficient, in spite of their potentially limited time and skills (Haefliger et al., 2008). Making the raw data widely accessible is also necessary because participants are scattered geographically and typically collaborate via web interfaces. Similar to the case of open source software, this high degree of openness may also be supported and reinforced by cultural norms and values shared in the crowd science community (see Mateos-Garcia & Steinmueller, 2008; Osterloh & Rota, 2007; Stewart & Gosain, 2006).

We suggest that these two dimensions – openness in participation and with respect to intermediate knowledge – do not only describe common features of crowd science projects, but also differentiate crowd science projects from other types of knowledge production. In particular, Figure 1

shows how differences along these two dimensions allow us to distinguish in an abstract way four different knowledge production regimes.

Figure 1: Knowledge production regimes with different degrees of openness



Crowd science is located in the top-right quadrant of Figure 1: projects are open to the contributions of many individuals and intermediate knowledge is openly disclosed. Note that this high degree of openness is also characteristic of open source software development, although the goal of the latter is not the production of scientific knowledge but of software artifacts.

The bottom-left quadrant captures projects that limit contributions to a relatively small and pre-defined set of individuals and do not openly disclose intermediate knowledge. 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 (see David, 2008; Murray & O'Mahony, 2007; Sauermann & Stephan, 2012). However, while traditional science is indeed "open" in that sense, it is largely closed with respect to the two dimensions captured in our framework. 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 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 works like a highly competitive system of tournaments, inducing scientists to expend great effort and to disclose research results as quickly as possible. However, competition also discourages scientists from helping competitors, explaining why data, heuristics, and problem solving strategies are often kept secret (Dasgupta & David, 1994; Haeussler et al., 2009; Walsh et al., 2005). Of course, even scientists working in traditional science sometimes share intermediate results at conferences or discuss them with trusted colleagues to receive feedback. However, this very limited (and often strategic) disclosure is qualitatively different from the structural openness characteristic of crowd science.

Recognizing that secrecy with respect to data and intermediate results 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.[10] As such, an increasing number of projects have moved from quadrant 4 – "traditional science" – into quadrant 3 (bottom right). Even though projects in this quadrant are more open by disclosing intermediate knowledge after the project is finished, data and logs are still kept secret for the duration of the project. Moreover, participation in a given project remains limited to a relatively small number of individuals, often with pre-existing personal ties.

Finally, projects in quadrant 1 (top-left) solicit contributions from a wide range of participants. However, intermediate knowledge is not publicly disclosed. Moreover, while not explicitly reflected in Figure 1, 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 menial tasks such as collecting data from websites) and – more interestingly – innovation contest platforms such as Kaggle or Innocentive. In innovation contests, inquirers post their problems online and offer monetary prizes for the best solutions. Anybody is free to compete by submitting solutions. Once a winner is determined, he or she will typically transfer the property right to the solution to the inquirer in return for the prize money (Jeppesen & Lakhani, 2010). Thus, the projects in quadrant 1 are open to a large pool of potential contributors, but they limit the disclosure of results and data. A main reason for the latter is that project sponsors are often private organizations that seek to gain

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

some sort of a competitive advantage by maintaining unique access to research results and new technologies.[11]

Overall, while openness with respect to project participation and with respect to the disclosure of intermediate results 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 some important qualitative differences between crowd science and other regimes of knowledge production. Before we discuss in more detail the implications of these particular aspects, we explore some important differences among crowd science projects.

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

We now turn to a discussion of two sources of heterogeneity among crowd science projects because this heterogeneity may have important implications for our understanding of the benefits and challenges crowd science projects may face. As reflected in Figure 2, these dimensions include the nature of the task as well as skill requirements.[12]

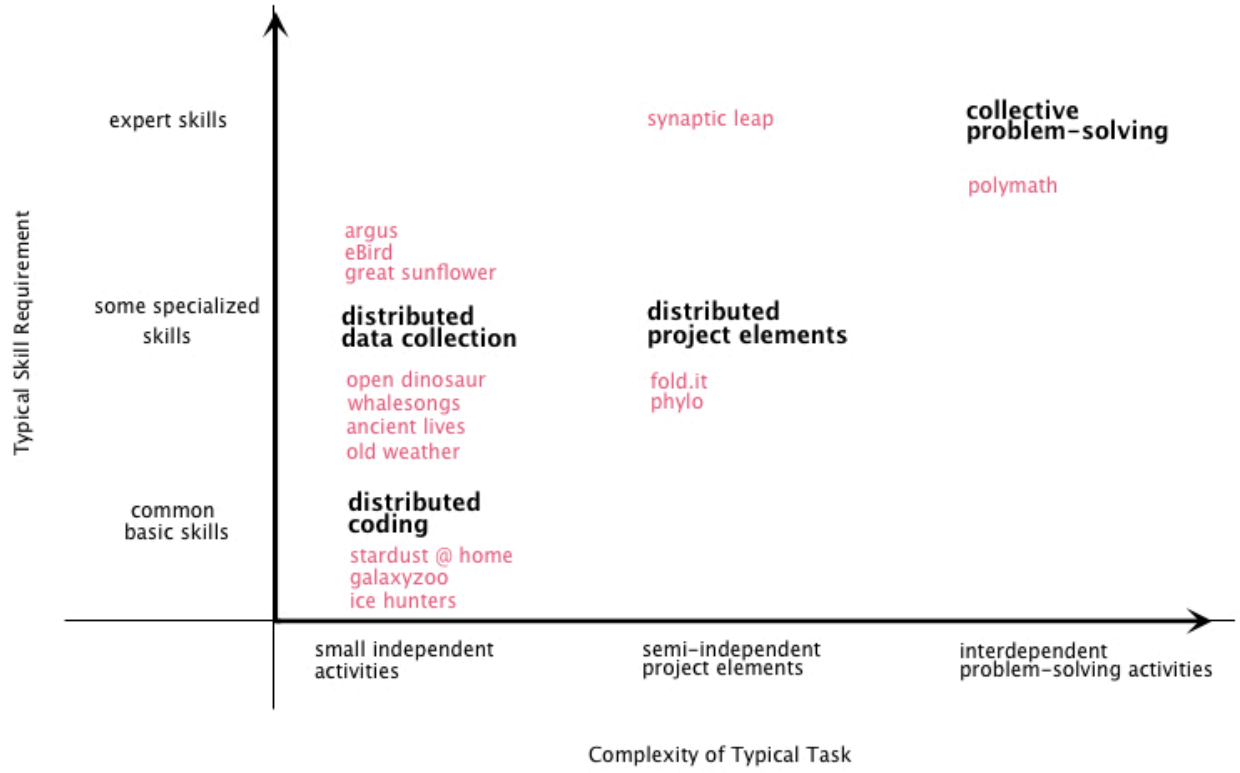
First, our examples showed that project participants engage in a broad range of different tasks, including the coding of astronomical data, the optimization of protein structures, or the collective development of complex mathematical proofs. While these tasks differ in many ways, it is particularly useful to consider differences in the complexity of the tasks carried out by individual contributors, where complexity refers to both the size of the task and to interdependencies with tasks performed by others (see Felin & Zenger, 2012). A large share of crowd science projects involves primarily tasks of low complexity. For example, the coding of images in Galaxy Zoo is a relatively simple task, with each contribution taking only a small amount of time and being independent of the activities of other contributors. Indeed, one may think that these 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 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, 2012).

[11] Innovation contests are sometimes considered as elements of broader "open innovation" strategies of firms (see Chesbrough, 2003; Dahlander & Gann, 2010; Felin & Zenger, 2012). While we will consider potential firm involvement in crowd science below, most existing crowd science projects do not involve firms.

[12] 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.

Projects located in the middle of the x-axis in Figure 2 involve more complex tasks, including some that require a considerable amount of creativity and human intelligence. For example, the Foldit platform used players not simply as "free labor" but explicitly studied how players approached problems and find solutions that were faster and better than those achievable with available computer algorithms. Based on these observations, in turn, Foldit was able to significantly improve subsequent computer algorithms. Finally, tasks with high levels of complexity include distributed problem-solving tasks that typically require participants to integrate contributions of other members in an interactive fashion and where each meaningful contribution can require a larger amount of effort and time.

Figure 2: Heterogeneity among crowd science projects



While our examples show that existing crowd science projects involve a wide range of tasks, Figure 2 also shows that most projects are located to the left of the figure, i.e., they primarily involve relatively simple tasks. One potential reason for the prevalence of simple tasks is that such tasks allow for the participation of a wider range of contributors, including "citizen scientists" without a formal training in science. This observation suggests that it is useful to consider in more detail the skills required by project participants. Many citizen science projects such as Galaxy Zoo ask for contributions that require no special skills. Complex tasks such as participation in Polymath problems typically require higher levels of

specialized skills, as evidenced by the prevalence of trained mathematicians among Polymath participants. Interestingly, the mapping between problem complexity and skill requirements is not completely linear. In particular, several projects involve relatively simple data collection tasks that still require specialized skills and knowledge of the particular subject matter. For example, consider eBird, which asks birders to submit data on the types of birds they see or hear in a particular location, requiring considerable skills in the identification of birds. Similarly, the Argus project relies on captains of ships to collect measures of seabed depth using sonar equipment. Finally, it is worth highlighting that the specialized skills required in projects are not necessarily acquired as part of formal scientific training. For example, success in protein folding requires the ability to visualize and manipulate three-dimensional shapes in space, a cognitive skill that has little to do with the traditional study of biochemistry.

4 Potential benefits of crowd science

In the prior section, 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 knowledge. These two characteristics, in turn, suggest potentially important advantages crowd science projects may have in producing scientific knowledge. In the following, we discuss four such advantages, while also considering that some types of projects may be more likely to realize these advantages than others.

Access to human resources. Many crowd science projects mobilize previously inaccessible human resources, allowing them to address more challenging problems and to speed up the process of discovery. As the Galaxy Zoo example illustrates, the increase in labor inputs allows projects to leverage scarce material resources such as telescopes but also the unique expertise of highly trained project leaders by complementing them with the help of thousands of volunteers. Could lead scientists rely on powerful computers to accomplish these tasks without the involvement of a larger number of people? While this may be possible in some cases, an intriguing aspect of crowd science is that contributions are typically uncompensated, significantly limiting the monetary cost.[13] Moreover, it turns out that humans continue to be more effective than computer algorithms in several realms of problem solving, including image and sound detection (Malone & Klein, 2007). Humans may also rely on intuition to improve optimization processes, whereas computer algorithms often rely on random trial-and-error 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

[13] This benefit is also very salient in OSS development. For example, once Linus Torvalds had posted his code in 1991, increasing numbers of programmers joined the effort, resulting in a tremendous growth of the Linux system. Moreover, much of the work was supplied for free by volunteer contributors – contributions which some calculations value at over 3 billion dollars (http://linuxcost.blogspot.com/2011/03/cost-of-linux.html; retrieved November 9, 2011).

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.

The foregoing discussion suggests that some types of projects may benefit more from access to the crowd than others. First, benefits should be particularly large for projects that involve a large number of tasks that require only skills that are relatively common in the general population, thus allowing a greater number of individuals to join the effort. As noted earlier, this rationale may explain why most existing crowd science projects are located in the bottom-left corner of Figure 2. Second, benefits should be large for projects that require distinctly "human" skills and where computers provide no viable alternative. Third, access to the crowd may also be particularly beneficial for projects that require rare skills, especially if those skills are not part of the standard scientific training. While such skills will likely be absent in any particular traditional project team, crowd science projects can "broadcast" the skill requirement to a large number of individuals and may thus be able find suitable contributors with the required skills (see Felin & Zenger, 2012; Jeppesen & Lakhani, 2010).

Knowledge diversity. Many problem-solving tasks require individuals to draw on prior knowledge and experience. As such, a diverse set of project participants provides more inputs that can be used in the development of new and creative solutions (Fleming, 2001; Hargadon & Sutton, 1997; Uzzi & Spiro, 2005). Crowd science can draw on a larger variety of knowledge simply due to a larger number of project contributors compared to traditional science projects. More importantly, these contributors often transcend organizational and field boundaries. Consider again Polymath, which attracted not only academic mathematicians, but also math teachers, students, and people from many other walks of life. That being said, not all projects will benefit from knowledge diversity to the same degree. We expect such benefits to be largest for collective problem solving tasks, but limited for tasks located to the left of Figure 2, i.e., data related tasks that are simple and require little creativity.

Some projects also benefit from diversity with respect to the geographic location of contributors, including projects to the left of Figure 2. In particular, efforts such as eBird, Argus, or the Great Sunflower Project seek to collect data from different locations to achieve comprehensive coverage of a large geographic space. The Great Sunflower Project, for example, asks participants to grow flowers that attract bees and count the number of bee visits that they see during rounds of 15-minutes daily observations. These kinds of projects require both a large number of participants as well as diversity regarding contributors' geographic location. As illustrated by this example, the benefits of geographic diversity are likely to accrue primarily to projects requiring observational data where geography itself plays a key role.

Verification. In traditional science, the verification of results occurs primarily through review by a small number of professional peers before results are published. While this traditional model has worked reasonably well, it is unable to detect all errors or cases of fraud (Lacetera & Zirulia, 2011). Replication and verification are particularly difficult in empirical work because investigators often do not disclose data or programs, and because the data themselves could be flawed (Dewald et al., 1986). As such, verification may often require a new data collection. The resulting high cost of verification imply that only radical discoveries will be scrutinized, whereas more incremental results are less likely to be checked (Lacetera & Zirulia, 2011). Higher degrees of openness in crowd science projects may improve verification for a number of reasons. In particular, a larger number of individuals who are actively involved in the project are more likely to come across inconsistencies, whether these inconsistencies result from error or mischievous intent. By way of example, this mechanism is salient in the Polymath project, where the relatively large number of "eyes" ensured that mistakes were quickly detected. Moreover, while peer review does not typically entail a detailed examination of the data or lab books, the openness of logs and data in crowd science projects allows observers to follow and verify the research process in considerable detail. In many cases, logs and other types of intermediate knowledge may also allow observers to tie claims to specific individuals, ensuring that contributors can be recognized for what they have accomplished, but also held responsible in cases of misconduct (see Rennie et al., 1997). Finally, the availability of a large pool of labor allows some projects to "replicate" tasks during the actual research process: they ask multiple contributors for identical contributions and compare the results to find inconsistencies. For example, in Galaxy Zoo 1, each galaxy was coded by up to 50 different persons.

Despite these potential advantages regarding the verification of results, some important caveats have to be kept in mind. First, many projects involve contributors who are not trained scientists and who may thus 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 (as exemplified by the repeated coding of imagines in Galaxy Zoo). Relatedly, even if the crowd science project discloses intermediate knowledge, the verification of some of that knowledge may require access to expensive physical resources that are not available to the preponderance of project participants or to outsiders. For example, considerable resources may be required to verify raw data such as the SDSS images posted on the Galaxy Zoo platform, or to verify complex statistical analysis performed on the coded Galaxy Zoo data. For these aspects of a project, crowd science may not offer any verification advantages compared to traditional science. Finally, it is possible that at least some of the non-professional contributors on a project care less about the quality of their work than would professional scientists, especially when they are only "trying out" whether they are a good fit with the

project. As such, automated mechanisms to identify and correct mistakes may be particularly important in projects involving large numbers of citizen scientists.

Accumulation of knowledge. A final benefit of openness emerges not at the level of a particular project but for the more general progress of science. Scholars of science have argued that the open access to research results 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 of final project results, additional benefits may accrue if other scientists can build on the intermediate knowledge 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. But even logs and discussion archives can be enormously helpful for future researchers if they provide insights into successful problem solving techniques. For example, the logs developed in the course of the Polymath project may be helpful for mathematicians as they seek to solve similar problems in the future.

5 Challenges and potential solutions

While openness with respect to project participation and intermediate knowledge may lead to considerable benefits, the same characteristics also create significant challenges. In addition to highlighting some of these challenges, we will draw on the broader organizational literature as well as insights from existing crowd science projects to point towards organizational and technical tools that may be useful in addressing them.

5.1 Organizational challenges

Matching projects and people. One key feature of crowd science projects is their openness to the contributions of a large number of individuals, many of whom may not know each other at the beginning of the project. Thus, organizational mechanisms are needed to allow for the efficient matching of projects and potential contributors. One potential approach entails systems that collect information on potential contributors and that allow project organizers to reach out to individuals with certain skillsets or research interests. Such a system exists already in Brazil, where the Lattes Platform provides a centralized database collecting research profiles, publications, and other information for virtually all scientists trained or working in that country.[14] An alternative approach makes it easier for potential contributors 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

[14] http://lattes.cnpq.br/; retrieved October 3, 2012.

find projects that fit their particular interests and levels of expertise. In that case, the matching occurs primarily due to the self-selection of individuals who may be completely unknown to project organizers when the project is initiated (see Felin & Zenger, 2012).

We expect that a particularly efficient way to solve the matching problem will be to bring projects and people together on larger hosting platforms that utilize a common infrastructure for multiple projects. Especially if these projects are similar with respect to the field of science, types of tasks, or skill requirements, they may benefit from a larger pool of potential project participants. Indeed, successful platforms 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 development[15] 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 in the areas of space, climate, humanities, and nature. When a new project is initiated, the Zooniverse platform routinely draws upon its large and growing base of existing contributors to recruit participants.

Division of labor. Most scientific research projects involve several different problem solving steps and exhibit a high degree of complexity. In order to allow a large number of contributors to work in a distributed fashion, these projects typically have to be divided into subproblems and tasks that can be assigned to individual participants. Open source software development has faced similar challenges and has developed sophisticated modularization techniques to overcome them. The basic idea of modularity is that a large problem can be 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 keeps the design of the overall project centralized and provides less skilled contributors with a small piece of the problem so that they can work independently and at their own pace. Of course, not all problems can be easily modularized, and projects with highly interdependent components are unlikely to be able to use as many participants as projects with a higher degree of modularization. Indeed, this logic may explain two interesting differences between our three case studies: Galaxy Zoo and Foldit both involve a much larger set of contributors, each of whom performs relatively simple tasks. Polymath, on the other hand involves a much smaller set of contributors, each of whom performs relatively complex tasks in interaction with other contributors. We suspect that mathematical problems are inherently less amenable to modularization, while empirical research offers more opportunities for a division of labor.

[15] Examples include sourceforge.com, Savannah and BerliOS.

Integration of contributions. Just as important as distributing tasks is the effective integration of the individuals' contributions to address the original research question. In highly modularized data collection and coding tasks (bottom-left of Figure 2), individual contributions can easily be integrated into larger data sets. Similarly, in some problem solving tasks, each contribution is a stand-alone solution and standard evaluation criteria can be used by the project leaders or by the community to judge the merits of the contributions (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, potentially deterring new entry (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.

Project leadership. Most crowd science projects require a significant amount of project leadership. Depending on the nature of the problem, the leader is fundamental in framing the scientific experiment, providing thoughtful modularization, securing access to 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 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 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. Finally, many projects require access to resources for the development of a technical infrastructure or for the acquisition of instruments and data. In existing crowd science projects, these resources have typically been provided by a small number of project leaders either from their "own" stock of resources (in the case of professional scientists) but also through the solicitation of project-specific government grants (see Årdal & Røttingen, 2012). In most existing crowd science projects, leadership positions seem to be held by professional scientists. In principle, however, 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 (some of the employees at Topcoder or Innocentive fulfill that function). 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 Motivation and incentives

The success of most crowd science projects depends on the degree to which they are able to attract a large and diverse set of contributors – including project leaders who fulfill important organizational functions. As such, a key question is how projects can motivate potential contributors to join and actively participate. Given the heterogeneity in potential contributors, it seems important to consider a wide range of intrinsic as well as extrinsic motives (Ryan & Deci, 2000; Von Krogh et al., forthcoming). The following discussion of motives for crowd science participation is not meant to be exhaustive. Rather, it seeks to illustrate the diversity of possible motives, to explore linkages between particular types of motives, project characteristics, and types of contributors, and to point towards mechanisms that may increase project participation.

5.2.1 Motivations to participate

Challenge, play and social interaction. Scholars have for a long time emphasized the important role of intrinsic motives, especially in the context of science and innovation (Amabile, 1996; Hertel et al., 2003; Sauermann & Cohen, 2010; Shapin, 2008). Intrinsically motivated people engage in an activity because they enjoy the intellectual challenge of a task or because they find it fun (Amabile, 1996; Ryan & Deci, 2000). Similarly, the feeling of accomplishment and competence created by successful task performance may be a powerful intrinsic motivator (Hars & Ou, 2002; Osterloh & Rota, 2007). Intrinsic motives are likely to play an important role in crowd science projects and may be particularly important for contributors who are not professional scientists and are thus unlikely to derive some of the extrinsic benefits discussed below. Intrinsic motivation may be easier to achieve for tasks that are inherently interesting and intellectually challenging, e.g., projects to the right of Figure 2. However, even simple and potentially tedious tasks – such as coding large amounts of data or systematically trying different configurations of molecules – can become intrinsically rewarding if they are embedded in a game-like environment. For example, many contributors enjoy playing the Foldit and Phylo games, and achieving an excellent score provides them with a feeling of pride and satisfaction (Nov et al., 2011). Not

surprisingly, an increasing number of crowd science projects seek to employ "gamification" to raise project participation (Prestopnik & Crowston, 2011).

Project participants may also feel good about being part of a particular project community and may enjoy personal interactions with other participants (Hars & Ou, 2002). While some types of projects – especially those involving collaborative problem solving – will naturally provide a locus of social interaction, projects can also actively stimulate interaction. For example, the Great Sunflower project seeks to find group leaders who operate as a local facilitator or collector of data in neighborhoods, communities or schools.[16] Other projects promote social interactions by providing dedicated 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.[17]

A particularly interesting aspect of intrinsic and social benefits is that they may be non-rival, i.e., that the benefits to one individual are not diminished simply because another individual also derives these benefits (Bonaccorsi & Rossi, 2003). Indeed, social benefits may even increase as the number of contributors increases, potentially leading to positive network effects.

Knowledge and understanding of particular topics. A second important motive for participation is an interest in the particular subject matter of the project. To wit, the motive most often mentioned by contributors to Galaxy Zoo was an explicit interest in astronomy (Raddick et al., 2009). 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. An especially powerful version of interest in a particular problem is evident in the growing number of projects 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 have a very strong personal stake in the success of the project, leading them to make significant time commitments (see Marcus, 2011). Recognizing the importance of individuals' interest in particular topics, some platforms such as Galaxy Zoo have enriched the work by providing scientific background information and interesting trivia to project participants. 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 "amusing" 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 those projects to identify a sufficient number of individuals with matching interests. More generally, this point reinforces

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

the importance of organizational mechanisms to facilitate the matching between projects and individuals with respect to both skills and interests (see section 5.1).

Recognition and career concerns. The motive to earn peer recognition has always been central in the institution of traditional science (Merton, 1973). In crowd science projects, this motive may thus be a powerful motivator especially for professional scientists whose careers depend on the production of scientific knowledge. These scientists may be particularly sensitive to the question how credit for project results is assigned. As such, it is interesting that some crowd science projects such as Galaxy Zoo, Phylo and Foldit assigned (co-)authorship to some of the individual contributors (typically the project organizers), while other projects instead include on the resulting publications only the group as a whole, such as in the case of "D. H. J. Polymath". While the use of group pseudonyms is consistent with the spirit of a collective effort, we suspect that it may reduce project participation by professional scientists who need publications to succeed in the traditional institution of science. Of course, participants may also gain recognition through mechanisms other than authorship. For example, projects that involve close collaboration among participants (i.e., those to the right of Figure 2) may also provide a unique opportunity for participants to build a reputation for creative contributions and to get noted by others.

Money. In the current landscape of crowd science projects, direct financial rewards for project participation are largely absent, at least for non-organizers. We suggest that 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. Moreover, for tasks that involve low levels of skill, the supply of contributors may be quite large, and a sufficient number of intrinsically motivated contributors may effectively push "wages" to zero (Brabham, 2008). Of course, we discussed earlier how a different kind of knowledge production regime – innovation contests – also draws on a large pool of contributors while employing financial incentives as its primary tool. A key difference is that participants in innovation contests typically deliver complete work products that can be attributed clearly to a particular participant and that can be judged and ranked with respect their performance. Conversely, crowd science projects often involve only relatively small modular contributions (each of which has a low value) or highly interactive collaborative contributions (where it is difficult to evaluate individual performance). More importantly, crowd science projects openly disclose both final project results as well as intermediate knowledge. As such, the overall project is unlikely to be able to generate significant financial returns (see Cohen et al., 2000), leaving little money to be distributed to contributors.[18] While this discussion suggests that monetary motives are likely to play a relatively small role compared to other

[18] Thus, while some tasks such as the coding of data could in principle be performed on a paid basis (e.g., using Amazon's Mechanical Turk), this approach is likely to be prohibitively expensive for science projects that plan to openly disclose intermediate knowledge and final data.

motives, financial motives may become more important in crowd science than they are now. In particular, consider that in open source software, an increasing share of programmers is paid by firms such as IBM or Intel, reflecting that firms have developed business models that allow them to benefit financially from open source software by developing proprietary value-added or by selling products that draw upon open source outputs (Hars & Ou, 2002; Lakhani & Wolf, 2006). It is conceivable that firms may find similar ways to benefit from crowd science, potentially leading to a more important role of financial motives for project participation. Indeed, an interesting example of a company-run crowd science project is Argus, which asks captains to collect data on ocean 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 maritime products.

5.2.2 Reconciling conflicting motivations

Crowd science projects typically involve a large number of individuals from diverse backgrounds, requiring us to consider potential conflicts between contributors with different motives and incentives (see Harhoff & Mayrhofer, 2010). While the extent and concrete nature of such conflicts remain ill-understood, the open source software experience suggests that different motivations can co-exist within a project, and that potential incentive conflicts can in principle be mitigated using contractual mechanisms. 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, the right to make derived works, and the right to impose further obligations on work derived from project results (McGowan, 2001). 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.[19] 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 would want 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

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

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 the first place. 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).[20] Finally, consider a fourth set of contributors – citizen scientists who simply enjoy watching videos of solar storms and 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 OSS 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, 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. 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.

The above example is meant primarily for illustration. While the open source experience suggests that well-defined and modular rights were central to making community-based production a stable phenomenon, systematic research is needed to examine whether and how contractual mechanisms can be usefully employed to mitigate incentive conflicts in crowd science projects. Despite their potential benefits, such contractual arrangements may also face various challenges. Among others, it is not clear how well license contracts can be enforced, although some crowd science projects such as Zooniverse already require contributors to log in and accept certain terms and conditions. In addition, some scholars suggest that narrowly-conceived rights may become too many and hinder the emergence of unexpected research trajectories (see Heller & Eisenberg, 1998; Murray & Stern, 2007). Given these potential concerns, future research on the costs and benefits of contractual arrangements in the specific context of crowd science is clearly warranted.

[20] 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.

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 science as well as on open source software development. Many of the conjectures developed in our discussion provide fertile ground for future qualitative and quantitative research. For example, research is needed on the degree to which modularization is currently being used – or can be employed – to facilitate the distributed work in crowd science projects. 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 citizen 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 levels of skill (see Figure 2). 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 2, most existing crowd science projects involve relatively simple tasks, many of which involve the generation 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 to projects that require different – and likely more complex – contributions. Examples such as Foldit and Polymath show that more complex contributions are feasible, yet we have also discussed some of the organizational challenges that are particularly salient in projects to the right of Figure 2. As such, future research on how crowd science project can be designed for a broader range of tasks and to efficiently tackle problems that are difficult to modularize would be of particularly great value.

Finally, future research is needed to examine firm involvement in crowd science projects, and which particular business models may prove profitable. As noted in section 5.2.2, it is conceivable that firm involvement requires hybrid approaches that preserve openness while still giving contributing firms preferential access to project outputs or allowing firms to appropriate value by complementing project outputs with firm specific resources and capabilities. Indeed, such approaches appear to develop in drug-related crowd science projects that recognize the need for industry partners to take on development functions (Årdal & Røttingen, 2012).

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 science. Similarly, crowd science projects that openly disclose discussions, problem solving attempts, and intermediate solutions may provide unique insights into the process of knowledge creation. For example, such detailed data may allow us to study the interactions among individuals in scientific teams (Singh & Fleming, 2010), to observe characteristics of both successful and unsuccessful problem solving attempts, and to trace back successful solutions to their origins. 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 some prominent examples of crowd science projects. We then developed a conceptual framework to distinguish crowd science from other regimes of knowledge production. In doing so, we highlighted two features of crowd science projects: openness with respect to project participation and openness with respect to intermediate knowledge. 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. In the final section of this paper, we will consider 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 knowledge, crowd science projects may provide greater "spillovers" to other projects and generate additional benefits for the general progress of science. 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 science 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 knowledge that is disclosed by crowd science projects but is 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.[21]

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 5.2.2, the open source software experience suggests that such tools can foster the development of community-based production[22] and may be particularly useful in reconciling potentially conflicting incentives of different types of project participants. Finally, funding agencies, policy makers, and scholarly organizations should engage in discussions regarding how the quality of research can be assured as participation in projects extends beyond professionally trained scientists and as the Internet offers opportunities to quickly disclose research results and data without the use of traditional journals and the associated process of peer review.

[21] http://blogs.loc.gov/digitalpreservation/2012/07/preserving-online-science-reflections/
[22] See for example the repository of the Open Source initiative, http://www.opensource.org/osd.html

REFERENCES

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.

Brabham, D. C. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. ***Convergence: The International Journal of Research into New Media Technologies***, 14(1): 75-90.

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

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***.

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).

Dewald, W. G., Thursby, J. G., & Anderson, R. G. 1986. Replication in empirical economics: The journal of money, credit and banking project. ***The American Economic Review***, 76(4): 587-603.

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. 2012. Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. ***Research Policy***.

Hargadon, A., & Sutton, R. I. 1997. Technology brokering and innovation in a product development firm. ***Administrative Science Quarterly***: 716-749.

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.

Heller, M., & Eisenberg, R. 1998. Can patents deter innovation? The anticommons in biomedical research. ***Science***, 280: 698-701.

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.

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.

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.

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.

Murray, F., & Stern, S. 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge? ***Journal of Economic Behavior and Organization***, 63: 648-687.

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

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

Nov, O., Arazy, O., & Anderson, D. 2011. Dusting for science: motivation and participation of digital citizen science volunteers. ***Proceedings of the 2011 iConference***: 68-74.

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

Osterloh, M., & Rota, S. 2007. Open source software development--Just another case of collective invention? ***Research Policy***, 36(2): 157-171.

Polymath, D. 2012. Deterministic Method to Find Primes. ***Mathematics of Computation***, 81(278): 1233-1246.
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.
Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. 2009. Galaxy Zoo: exploring the motivations of citizen science volunteers. ***arXiv preprint arXiv:0909.2925***.
Rennie, D., Yank, V., & Emanuel, L. 1997. When authorship fails: A proposal to make contributors accountable. ***Journal of the American Medical Association***, 278: 579-580.
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. 2012. 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.
Shapin, S. 2008. ***The Scientific Life: A Moral History of a Late Modern Vocation***: University of Chicago Press.
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.
Stewart, K. J., & Gosain, S. 2006. The impact of ideology on effectiveness in open source software development teams. ***Management Information Systems Quarterly***, 30(2): 291.
Uzzi, B., & Spiro, J. 2005. Collaboration and creativity: The small world problem. ***American Journal of Sociology***, 111(2): 447-504.
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.
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