Hirsh, Haym, "How Do You Cite a Crowd?"

How Do You Cite a Crowd?
A White Paper for NSF's "Changing the Conduct of Science in the Information Age" Workshop
November 12, 2010

Haym Hirsh
Rutgers University

- Three months ago *Nature* published an article concerning Foldit, a computer game whose top players are non-scientists who beat the best protein structure prediction programs.
- In 2009 a new proof of the density Hales-Jewett theorem, the first to use elementary methods, was jointly crafted by more than three dozen participants via social media, described in a paper whose author is given as D.H.J. Polymath.
- That same year researchers at Google and the Centers for Disease Control (CDC) published a paper in *Nature* showed that tracking frequencies of flu-related Google queries allows detection of flu outbreaks over a week earlier than the CDC.
- Also that year, *Current Biology* published a paper that showed that species that exhibit vocal mimicry also exhibit motor entrainment to music – they move to the music's rhythm – in part by analyzing YouTube videos of animals.
- In 2006 *Nature* published a paper on "The scaling laws of human travel" that used records from wheresgeorge.com, a website at which people can enter and track currency they have possessed.
- Tens of thousands of people have used Galaxy Zoo to classify over 40 million astronomical objects, leadings to such discoveries as the fact that neighboring galaxies have aligned spin directions.
- Computer users around the world allow their machines to be networked into large distributed supercomputers that compute prime numbers, compute protein folding, break encryption systems, and search for signals of extraterrestrial life, among many others.
- The computational linguistics and computer vision communities, which rely heavily on machine learning over corpora of data, increasingly use Amazon Mechanical Turk to micro-outsource the human labor of data labeling.
- Researchers at Stanford have shown how comparing not just biological sequences but also their associated literatures can improve homology search, and how large databases of structured knowledge can be populated by the pharmagenomic knowledge embedded in the relevant science literature.
- Researchers in computer vision and computer graphics are taking the billions of photos in community photo collections such as Flickr to construct rich, navigable, 3D depictions of the world they represent, and to cut out your ex-wife from a photo and splice in instead new content the seamlessly matches the rest of the photo.

Information and communication technology innovations are bringing people together in ways that have never previously been possible or even imagined. The area of collective intelligence seeks to understand these new ways in which people collaborate and create outcomes that are integrally about large groups of participating individuals, as much as they are about the new technologies that underlie them. As with the rest of our society, science must confront the challenges and implications of collective intelligence in the practice and communication of our scholarly work.

Who gets credit when the knowledge work that allows us to discover that neighboring galaxies have aligned spin direction comes from tens of thousands of individuals? What is the authorship of a paper when the ideas underlying a proof are distributed across a blog and over a thousand comments, especially when the authors themselves choose to use a pseudonym? How do we support repurposing of data so that we can discover airline travel patterns from a dollar bill tracking website, 3D structures from community photo collections, flu outbreaks from search engine queries, or correlation between motor entrainment and vocal mimicry from YouTube videos? Who gets credit if a new biochemical discovery is made by a non-scientist playing a game? How do we mine the scientific literature to discover the hidden wisdom that may span hundreds or thousands or more papers, where each paper contributes to the collective knowledge?

The question of attribution and credit is harder than we thought it was when we consider the new affordances for science of collective intelligence. It's not just about the new forms of data-intensive science that technology has enabled, where we may seek scholarly acknowledgement of such activities as data production, data stewardship, software development, and the like. It's not just about new forms of scholarly communication and peer review that are unlike what science has relied on for hundreds of years. The very nature of how people come together to generate new knowledge and new outcomes has changed, in ways that are incompatible with our established ways of viewing the science enterprise – whether digital or otherwise.

Science funding agencies can respond to these forces in a number of ways. The first is that whatever steps an agency takes to be effective stewards of science funding, they must be part of an ongoing process that can adapt to the increasingly fast-changing landscape of science. The second is to keep in the cross-hairs of all decisions the gold standard of science: Reproducibility. Thus, while data management plans provide a crucial element for reproducibility, they are a means to an end and not the end itself. Funding agencies can take steps to maintain a focus on reproducibility, such as by having reviewers explicitly comment on and assess the reproducibility characteristics of proposed projects. Third, funding agencies should be vigilant in supporting new modalities of science, and not themselves fall into set ways that reflect only older ways of conducting science. Finally, funding agencies must continue to be stewards of science cyberinfrastructure, keeping timely with what is necessary to support the changing landscape of science, lest we only support old ways of doing new science.