Final Progress Report
Sustainability Science Program, Harvard University
Term: September 1, 2011 – August 31, 2012
Submitted: July 2012

Name: Gabriel Chan

Your field(s):
Public Policy (economics, statistics, political science)

Your degree program, institution and (expected) graduation date:
PhD in Public Policy from the Harvard Kennedy School, expected 2014-2015

Faculty host(s) at Harvard name and department:
Bill Clark (HKS)
Venkatesh Narayanamurti (HKS, SEAS)

Description of SSP-related research activity:

Title: Evaluating Innovation at the National Labs: Automated Content Analysis of Patents and Matching

Abstract: The research project that I am conducting during my SSP fellowship seeks to evaluate energy and environmental innovation at the U.S. National Labs. My project uses patent data from energy and environmental technologies to estimate the effect of an innovating institution on subsequent innovation and technology deployment. Evaluating an institution's innovation effort is made difficult because the research scope of institutions often have partial but not complete overlap with each other, implying that innovation arising from one institution can be compared to only a very carefully selected subset of other innovations. The key challenge of this research will be to identify the most appropriate patents for comparison so that the differences in the effect of institutions on subsequent innovation can be estimated, holding the differences in the technological scope of the patents constant. This project applies a natural language processing algorithm, the latent Dirichlet allocation topic model, to the corpora of U.S. patent abstracts to estimate the effect that the U.S. National Labs have on subsequent innovation as compared to innovations from private sector institutions. Utilizing a matching algorithm on the modeled topic structure of patent abstracts, this paper identifies an appropriate subset of patents filed by the private sector that can be compared to patents filed by the National Labs. Then, subsequent citation rates between public and private sector patents are compared, holding the differences in the technological scope of the patents constant.

For policymakers considering privatizing public National Lab R&D effort, this is the most relevant metric for estimating the counterfactual outcome that would result if the same R&D that was conducted in a national lab was instead conducted by the private sector. My research will develop and apply cutting-edge social science statistical methods to identify and estimate causal relationships in the observed patenting behavior of the National Labs to develop empirically-grounded policy recommendations for energy and environmental R&D management decision-making. This project is also one of the first research efforts to combine natural language processing methods with matching methods in an applied context in the social sciences.

In future stages of this work, I will use the terms of the licensing agreements that the National Labs enter into with the private sector to commercialize technology that the Labs develop. This will allow me to observe a monetary value for a patent or group of patents. By combining the topic modeling approach with data from licensing agreements, I hope to estimate how shifts in National Lab funding have affected the value of Lab R&D in specific technological areas.

Identification of the problem you address:
Innovation of new technologies, particularly innovation of technologies which supply a public good, is often seen as a necessary component of a transition towards sustainability. The innovation of new technologies occurs through the advancing frontier of scientific and technological possibility within the context of institutional arrangements that support science and technology development and deployment. While policy has evolved over the last decades to engage public resources in the innovation process, little empirical evidence has been brought to bear on how decisions should be made with respect to the institutional arrangements that best facilitate publically-guided innovation effort. The problem that I am addressing in my SSP fellowship compliments the work being conducted in the Harvard Innovation for Sustainability Working Group.

Key question asked about the problem:
How effective are institutional arrangements such as the U.S. National Labs in spurring subsequent innovation and commercialization of energy and environmental technologies for sustainability?

The methods by which you answered that question:
I formulate my question of interest in the Rubin Causal Model framework (Rubin 2006) to clearly estimate the causal effects of interest that are identifiable in the data. In this framework, the units of analysis are *patentable potential but as yet undiscovered* innovations. This description of the innovation process posits a fixed (at any point in time) distribution of potential innovations which innovators draw from. The units of analysis receive one of several possible observed treatments: being discovered and patented in a National Lab/university/corporate R&D lab/etc. Potential innovations which are either not discovered or not patented are not observed. Finally, outcomes are the citations that a patented innovation subsequently receives. Citations are not a perfect outcome metric, in part due to the potential spurious relationship between treatment assignment and citations (e.g. an innovator may find more prestige in citing a national lab patent). However, I will use citations for this project as an initial exploratory outcome variable due to the lack of other easy to quantify and interpret metrics. As a next step in this project, I could use a more nuanced outcome metric such as subsequent citations by firms other than the firm that filed the cited patent (to capture commercialization spillovers), or citations by industries other than the industry that filed the cited patent (to capture inter-industry spillovers). This setup makes it immediately clear that treatment is assigned non-randomly. Because institutions innovate in different technological spaces due to their distinct incentive schemes, access to human and capital resources, and the unique tacit skills they can draw upon, it would be unrealistic to assume that a potential innovation, conditional on it having been patented, would have been equally likely to have been discovered and patented in every institution. For this project, I will move beyond making direct comparisons, by applying matching methods to create a balanced sample of patents from different institutional arrangements. I will create these balanced samples by first estimating the individual-level textual structure of patent abstracts by using natural language processing methods. This structure, which can be estimated at any specified granularity, can then be used to estimate predicted frequencies of structural elements of the text, referred to as "topics," which can be balanced just as any other covariates are balanced in matching studies. With this empirical strategy, I will be able to find a group of patents from one institution which "match" patents from a different institution on the substantive content of the patent, allowing me to control for the extensive differences in the innovation direction of two institutions. This will then allow me to estimate differences in the *intensive* innovative activity of two institutions controlling for their research scope. For decision-makers who seek to fund directed R&D (e.g. in a particular technological area), the intensive difference in innovative activity is the relevant metric for comparison since the extensive margin of innovation is often fixed in practice.

Principle literature upon which the research drew:
The methodological literature that I draw upon has two components: 1) matching, primarily the work of Don Rubin, and 2) natural language processing topic modeling, primarily the work of David Blei. The substantive literature that I draw on has its roots in innovation economics and policy. In particular, the empirical studies in innovation economics using patent data has its roots in the works of Schmookler, Griliches, Scherer, Pakes, Hall, Jaffe, and Trajtenberg.

Empirical data acquisition description:
The main data source for this project is the text contained in patent abstracts from patents on technologies developed in the national labs and in the private sector. Full plain-text formatted patents are available in both HTML (from the US Patent and Trademark Office) and XML (through Google). One particular challenge of this project will be identifying which patents were developed through national lab research. Jaffe and Lerner (2001) used several databases from the U.S. Department of Energy to identify National Lab patents in previous research, but I have received a comprehensive database of National Lab patents from the Department of Energy's Energy Innovation Portal.

Geographical region studied:
I will begin my research using U.S. patent data. Patent data from outside the U.S. is also available (e.g. from the EPO and WIPO) and could give an international dimension to my research.

Recommendations that might be relevant for your problem:
My research will develop recommendations that may be relevant for policymakers seeking to allocate directed R&D effort in environmental and energy technologies across public and private research institutions.

A description of the final product(s) you have/are aiming to produce; Description of major other intellectual or professional advancement activity(ies) over the past academic year:

1. I passed my PhD qualifying exam in June 2011 and will defend my dissertation prospectus, based on my SSP research project, in August 2012.

2. I have been accepted to present at the 2nd Global TechMining Conference at the 17th International Conference on Science and Technology Indicators, Université du Québec à Montréal, Canada in September 2012 with the possibility of publishing my work in the journal *Scientometrics*.

3. I am a researcher in the Sustainability Science Initiative on Innovation for Sustainable Development in the Energy Sub-Group. I am organizing research on case studies of carbon capture and sequestration technology on of clean cookstoves.

4. I am a Chapter Scientist for the 5th Assessment Report of Working Group III of the Intergovernmental Panel on Climate Change (IPCC). In this capacity I am surveying the literature on international climate policy architectures and mechanisms for technology development, transfer and diffusion.

5. In addition to the research projects that I am conducting independently for my SSP Fellowship, I am also collaborating with a team of researchers on a project to evaluate the effectiveness of U.S. state renewable energy portfolio policies and on a separate project to evaluate the effectiveness of the Clean Development Mechanism.

Citations for reports, papers, publications and presentations that built on your fellowship research:

1. **Chan, Gabriel,** Robert Stavins, Robert Stowe, Richard Sweeney. 2012. The SO2 Allowance-Trading System and the Clean Air Act Amendments of 1990: Reflections on 20 Years of Policy Innovation. *National Tax Journal*, 65: 419-452; http://ntj.tax.org.

The introduction of the U.S. SO_2 allowance-trading program to address the threat of acid rain as part of the Clean Air Act Amendments of 1990 is a landmark event in the history of environmental regulation. The program was a great success by almost all measures. This paper, which draws upon a research workshop and a policy roundtable held at Harvard in May 2011, investigates critically the design, enactment, implementation, performance, and implications of this path-breaking application of economic thinking to environmental regulation. Ironically, cap and trade seems especially well suited to addressing the problem of climate change, in that emitted greenhouse gases are evenly distributed throughout the world's *atmosphere. Recent hostility toward cap and trade in debates about U.S. climate legislation may reflect the broader political environment of the climate debate more than the substantive merits of market-based regulation.*

2. Jenner, Steffen, **Gabriel Chan**, Rofl Frankenberger, Mathias Gabel. 2012. "What Drives States to Support Renewable Energy?" *Energy Journal*, 33(2). http://ideas.repec.org/a/aen/journl/33-2-01.html

Why do states support electricity generation from renewable energy sources? Lyon/ Yin (2010), Chandler (2009), and Huang et al. (2007) have answered this question for the adoption of renewable portfolio standards (RPS) at the U.S. state level. This article supplements their work by testing the core hypotheses on the EU27 sample between 1990 and 2010. Furthermore, the article asks why the majority of EU states relies on feed-in-tariffs (FIT). The study conducts logistic time series cross-section regression analyses that run on a hazard model. Evidence in support of private interest theory and public interest theory is provided. (a) The existence of a solar energy association increases the probability of a state to adopt regulation. (b) Solar radiation, and (c) the unemployment rate also increase the odds. (d) Electricity market concentration decreases the probability of transition.

3. Anadon, Laura Diaz, Matthew Bunn, **Gabriel Chan**, Melissa Chan, Charles Jones, Ruud Kempener, Audrey Lee, and Venkatesh Narayanamurti. 2011. *Transforming U.S. Energy Innovation: Harvard Energy Research Development, Deployment, and Demonstration Report.* Cambridge, Mass.: Report for Energy Technology Innovation Policy research group, Belfer Center for Science and International Affairs, Harvard Kennedy School. http://belfercenter.ksg.harvard.edu/publication/21528

The United States needs a revolution in energy technology innovation to meet the profound economic, environmental, and national security challenges that energy poses in the 21st century. If the U.S. government does not act now to improve the conditions for innovation in energy, even in times of budget stringency, it risks losing leadership in one of the key global industries of the future, and the world risks being unable to safely mitigate climate change and to reduce vulnerability to disruptions and conflicts—both domestic and international. Waiting is not an option.

Researchers at Harvard Kennedy School undertook a three-year project to develop actionable recommendations for transforming the U.S. energy innovation system. We surveyed over 100 experts across a broad range of energy technologies; conducted extensive economic modeling; and developed and implemented a new methodology for assessing how much research, development, demonstration (RD&D) investment is needed, and in which technologies. This work also included: interviews with a range of energy innovators and policymakers; the first survey of energy innovation in U.S. businesses; analyses of how effectively the Department of Energy (DOE) interacts with private firms; case studies of the operations and effectiveness of key energy innovation institutions; and development of new data on international energy RD&D spending and cooperation. This research has led us to five key recommendations for accelerating U.S. energy innovation.

Principal collaborators outside Harvard (list name and institution):
Steffen Jenner, University of Tuebingen, Germany

Awards or grants that you have received this year for the current or coming year:
HKS Belfer Center Science, Technology, and Public Policy Fellowship for the 2012-13 academic year.