KAUFFMAN DISSERTATION EXECUTIVE SUMMARY

KAUFFMAN
ENTREPRENEURSHIP SCHOLARS

KAUFFMAN DISSERTATION FELLOWSHIP

Part of the Ewing Marion Kauffman Foundation's Entrepreneurship Scholars initiative, the Kauffman Dissertation Fellowship recognizes exceptional doctoral students and their universities. The annual program awards Dissertation Fellowship grants to Ph.D., D.B.A., or other doctoral students at accredited U.S. universities to support dissertations in the area of entrepreneurship.

Since its establishment in 2003, this program has helped to launch world-class scholars into the exciting and emerging field of entrepreneurship research, thus laying a foundation for future scientific advancement. The findings generated by this effort will be translated into knowledge with immediate application for policymakers, educators, service providers, and entrepreneurs as well as high-quality academic research.

Ewing Marion
KAUFFMAN
Foundation

www.kauffman.org/kdf

Dissertation Summary

Title: Essays on the Impact of Digital Information on Innovation

Date: June, 2016

Abhishek Nagaraj

nagaraj@mit.edu

Massachusetts Institute of Technology, Sloan School

Abstract:

This dissertation consists of three essays studying the impact of new, digital information on innovation in different markets. In these essays, I explore how new information about physical, natural and social phenomena shape the rate and direction of innovation. I use the metaphor of map-making to understand the role of information and argue that in much the same way that maps shape decision-making in uncertain environments, digital information shapes action in innovative and creative contexts. Further, similar to ways in which map-makers make important representational choices when constructing a version of "reality", information providers makes choices that shape the view of the landscape upon which entrepreneurs make decisions. Therefore, the very process through which information gets collected, constructed and represented becomes the object of study, and informational choices become determinants of variations in innovation and entrepreneurship outcomes.

In the three essays in this dissertation, I explore organizational, technical, legal and social processes that cause variation in the way that information is collected and represented, and show that these processes have a large and significant impact upon innovative outcomes. The first essay studies the process of digitization of archival magazines, and shows that intellectual property surrounding digital material influences the extent of diffusion and reuse on creative

activity on Wikipedia. The second essay studies Landsat, a NASA program to collect satellite images of the surface of the earth, and shows that idiosyncratic and unintended variations in the quality of the mapping effort affected the rate and direction of new discoveries and the success of entrepreneurial firms in the gold exploration industry. And finally, in the third essay I study the US Census TIGER mapping project, which aimed to generate a new street-map of the US, and show that the sequence in which information was collected affected community mobilization and new knowledge-creation on OpenStreetMap, an online street-mapping community.

Collectively, these three essays describe an agenda for understanding the mechanisms through which information providers could act as map-makers and shape innovative activity. Ultimately, "there is no such thing as a complete map", so how do incomplete maps come to be, how do they affect innovation and how should we design maps to yield the "best" outcomes? These are the questions that this dissertation hopes to inspire. In the rest of this document, I will discuss each of the three essays in turn, describing their research question, methodology and key results.

Essay 1: Does Copyright Affect Reuse? Evidence from the Google Books Digitization Project

The increasing digital representation of information has touched a wide breath of economic activities. Digitization has reduced the cost of storage, computation and transmission of information and enabled massive changes in the ways that creative producers build upon and reuse existing information. Despite its enormous potential, the digitization process is governed by intellectual property and copyright laws that were originally conceptualized for more traditional forms of content. Therefore, the question of whether and how copyright should be modified for the digital age has become prominent in policy and legal circles (Mergeset al.,

2012). Some firms have argued for strengthening of copyright protection given the difficulties in enforcing copyright on digital information (Anderson, 2007), while others have argued that the current copyright regime severely undermines reuse and therefore limits the economic potential of digitization (Samuelson, 1999; Lessig, 2004). Despite the prominence of these debates (for e.g. see Supreme Court case Authors Guild v. Google, Inc.), there is little empirical evidence about whether and how copyright influences the diffusion and reuse of digital information. A recent essay describing gaps in the literature summarizes this problem problem quite succinctly "what would be the economic effects of various alternative copyright arrangements and proposals for its redesign?"

In this essay, I make empirical progress on the question of the impact of copyright on the reuse of digital information by exploiting a natural experiment that occurred during a marquee project in the history of the internet: the digitization of about 30 million works by Google Books. While the project was still underway, in December 2008, Google Books digitized all existing issues of Baseball Digest, a prominent baseball magazine, and made them available online to readers for free. Apart from the fact that it is perhaps one of the most important reference sources on the game of baseball, Baseball Digest is particularly useful for my purposes because, due to an accidental failure to renew copyrights, issues of the magazine published before 1964 lapsed into the public domain. Consequently, pre-1964 Baseball Digest issues can be freely reused, while those published after 1964 are copyrighted and their reuse, without permission is legally prohibited. I focus on the impact of copyright on affecting the reuse of magazine material on Wikipedia. Not only is Wikipedia the fifth most visited website on the internet (receiving about 10 billion page-views every month) as well as a common source of information about the history of baseball, it also stores all past versions of a given page, allowing the analyst to track how information changes in response to the Google Books digitization event. Specifically, I track the reuse of information from Google Books on Wikipedia before and after the digitization event,

separately for the pre-1964 and post-1964 issues of Baseball Digest to understand how copyright affects the diffusion of digital material.

My data indicates that after the digitization of Baseball Digest in late 2008, the average number of citations for all publication-years of the magazine between 1944-84 increased dramatically, suggesting the large and positive impact of digitization on information reuse. However, when this increase is examined separately for in-copyright and out-of-copyright publication years, the gains from digitization are heavily concentrated for out-of-copyright issues. The econometric estimates indicate that citations to out-of-copyright publication years increase by about 135% as compared to citations to in-copyright publication years after digitization, even after controlling for a variety of different confounding factors. Further, I also consider the differential effect of the copyright law on affecting different types of pages and content on Wikipedia. I find that copyright mainly prevents the reuse of rare photos and images from Baseball Digest, but does not affect the reuse of textual material. The losses due to copyright are least salient for Wikipedia pages belonging to “superstar” baseball players because such players benefit from many other alternate sources of information. Copyright law therefore increases the inequality of information quality between more prominent and less prominent topics on Wikipedia.

This research contributes to literature on digitization and intellectual property in digital settings. I show, for the first time, how copyright law could severely curtail potential benefits of digitization in online contexts. Further, I also add to research on the role of digitization in influencing the differences in outcomes between more and less established players in a market. Finally, I also contribute to the nascent empirical literature on copyright including some work in the legal domain estimating the impact of copyright in the publishing context and work that studies the impact of copyright on prices, copyright enforcement and piracy.

Essay 2: The Private Impact of Public Maps--Landsat Satellite Imagery and Gold Exploration

Fundamental and basic knowledge about the physical world has led to new discoveries and massive increases in human prosperity since the middle ages. Economic history indicates that an important channel through which basic knowledge about the physical world could have enabled discovery is through novel maps of poorly understood geographies. For example, the Itenerario, a compendium of maps published in 1596 by the merchant Jan Huyghen Van Linschoten, contained basic knowledge about the East Indies including very delicate nautical data that provided insight into the currents, deeps, islands and sandbanks of unprecedented accuracy for those days" . Soon after this new map was published, the Dutch and British East India companies were established, many new territories and trading partners were discovered and the Portuguese monopoly over the trade and colonization in south-asia was ended. This anecdote begs the question: how does the arrival of basic knowledge through the publication of new maps causally affect the discovery of new opportunities and entrepreneurship in the private sector? This paper investigates this question in a modern context: the role of the NASA Landsat satellite mapping program in shaping the discovery of new deposits in the gold exploration industry.

Despite its status as one of the oldest forms of basic knowledge, the possible role of mapping information in shaping the geography of private discovery and entrepreneurship has resisted formal investigation. This is surprising because, unlike Borges' fantasy from the quote above, it is a cartographic truism that “there is no such thing as a complete map". In practice, even after a region has been mapped, it is quite common for many territories to have been ignored or poorly understood. While this variation in the availability of basic geographic knowledge across regions is quite prevalent, whether and how it affects the level and distribution of performance between larger and smaller firms is currently unknown.

In this paper, I propose that by opening the “black box” of mapping as an economic activity, it is possible to understand the role of public investments in basic knowledge on both industry performance and entrepreneurship. Specifically, I study the Landsat satellite mapping project and its role in shaping the discovery of new deposits in the gold exploration industry. Landsat provided the first images of Earth from space, and while the program was designed for its agricultural (and not geological) applications, maps from the program provided information that was relevant to guide early-stage gold exploration. Further, the Landsat program is a particularly appropriate setting because it represents a natural experiment with plausibly exogenous allocation of mapping information to some regions and not others. Specifically, while Landsat was designed to map the entire surface of the earth, in practice, there was significant variation in the timing of the mapping effort across different regions. Of the 9493 “blocks” (regions of 100 sq. mile each) which are needed for full coverage of the earth, some blocks received satellite maps early in the program, while others were mapped at significantly later points in time over the next decade. Further, quantitative assessments and qualitative interviews indicate that significant differences in the timing of the mapping effort were unintentional on the part of the program administrators, due to reasons like technical failures in satellite operation and cloud-cover in imagery.

The results suggest that, despite strong private incentives for mapping, the public Landsat mapping effort had a significant impact on the gold exploration industry. In baseline estimates, mapped regions were almost twice as likely to report a discovery when compared to unmapped regions. These differences imply meaningful benefits of the mapping effort in dollar terms---using rough estimates of discovery value (derived from data on the size of discoveries) the Landsat program led to a gain of approximately $17 million dollars for every mapped block over a fifteen year time period. For a country the size of the United States, this translates to

additional gold reserves worth about $10 billion USD that can be attributed to the information from the Landsat program. Having found that the Landsat program had large and positive benefits in terms of overall levels of discovery, I then turn to analyzing how these gains were distributed between different kinds of market participants. Specifically, I test whether the mapping program disproportionately benefited “juniors”, smaller and entrepreneurial firms in the exploration industry as compared to “seniors,” larger and more established players. I find that the smaller firms share an increased proportion of the new discoveries attributed to the Landsat program as compared to before the launch of Landsat. Specifically, while juniors were making about one of every ten new discoveries before the launch of the Landsat program, in blocks that benefit from the mapping program, they report one out of every four new discoveries, a considerable increase. Put differently, junior-led discoveries increased by a factor of 5.8, while the corresponding rise for seniors was only about 1.7, indicating that smaller firms benefited more than three times as much as incumbents from new mapping information. These results suggest that mapping information both raises the overall level of industry performance and disproportionately encourages the performances of smaller firms.

This paper contributes to the literature on the role of public investments in knowledge goods on encouraging the performance of firms in the private sector. While this literature is fairly extensive, this study joins some recent work evaluating this question using exogenous changes in the level of public investments on private patenting. The present study adds to this literature by highlighting a novel channel through which openly-available knowledge could matter for industry (investments in mapping goods), and by focusing on a direct measure of firm performance (discovery), rather than intermediate measures of performance such as patenting. This is also the first paper, to my knowledge, that finds that public information could differentially affect the performance of larger and smaller firms and thereby encourage entrepreneurship.

Essay 3: Does Open Data Spur Online Communities? Evidence from Crowdsourced Mapping

There is a firm belief that open access to government data can boost innovation and entrepreneurship. Guided by this belief, a number of organizations around the world have rallied around the idea of "open government data". The Obama administration passed an executive order in May, 2013 that "makes open the new default" for government information with the idea that such a policy would drive innovation and promote "the social good". In parallel with the movement for public access to government data, the role of online communities in driving innovation has been growing dramatically, based in-part on the mass adoption of the internet and reduced costs of collaboration. In particular "peer-produced" knowledge goods like open-source software, Wikipedia and crowd-science have been shown to contribute significantly to national growth and productivity.

Motivated by these two phenomena, in this project I analyze the role of public investments in new knowledge on spurring the development of community-based knowledge production. While there is little empirical work on this topic, proponents of open-data policies have argued that the availability of open data might spur knowledge production in online communities. A variety of different theoretical arguments could be employed to support this conclusion. For example, public investments in open data might reduce the cost of contributing new information, or attract new users by increasing the value of the platform.

In this project, I provide, the first empirical estimates of the role of public investments in open data on community-based knowledge production and show that open data might hinder rather than help the development of online communities. Specifically, I evaluate the impact of the US Census MAF/TIGER Improvement Project (MTAIP), a 8 year long, $200 million dollar project

undertaken by the US Census Bureau to provide a street-map database of the United States. I focus on the role of this open data project on encouraging the development of user communities on OpenStreetMap, a Wikipedia-style online mapping platform used widely across the internet in applications such as Craigslist, Foursquare and Apple Maps. I exploit a natural experiment caused due to a little known but highly consequential error made during the introduction of TIGER data in OpenStreetMap. When OpenStreetMap was launched in the United States, the community realized that they could leverage data on the street network of the US from the TIGER project launched by the Census Bureau. Accordingly, in 2006, computer programs were written that automatically incorporated all of the maps from the TIGER project within OpenStreetMap. However, the OpenStreetMap community did not realize that the TIGER project was slated to be completed in 2008, and the 2006 version of the map contained accurate information for only about 60% of the approximately 3100 counties in the United States, while the other counties contained incomplete and out-of-date information. Therefore a little more that half of the counties on OpenStreetMap benefited from public investments in the TIGER project, while the others did not. Further, my quantitative and qualitative analysis shows that counties that were included were very similar to the counties that were not, making OpenStreetMap a compelling natural laboratory to understand the impact of public investments in open data on crowdsourcing activity. I collect highly detailed data on about 1.3 million unique contributions to OpenStreetMap in the United States from 2006-2014 and link these contributions to regions that both received and did not receive TIGER data. By comparing blocks that were or were not affected by TIGER, before and after the inclusion of the public data, I am able to estimate the causal impact of public investments in bsaic knowledge on peer production in this setting.

The econometric results suggest that, despite the prevailing assumption that public sector investments in open data will boost peer production, the TIGER project seems to hinder rather

than encourage the development of the OpenStreetMap community. Specifically, counties that had received public investment from the TIGER project, had about 12% lower contributions and 6% lower user activity as compared to non-TIGER counties. Further, these differences seem to also affect follow-on knowledge production that was not influenced by the TIGER project. Specifically, meta-knowledge about the street network (such as speed limits, lane information, turn restrictions etc) which are vital for a usable digital map and which was never a part of the TIGER project, seems to be provided at a greater rate in non-TIGER counties as compared to TIGER counties. This suggests that lower user activity on OpenStreetMap in TIGER counties is not simply due to the fact that these counties have less missing information in the street layer -- the follow-on layers that never benefited from the TIGER information also seem to be affected.

These results speak to the literature on the role of the public sector in encouraging innovation in the private sector. While in practice, the role of open source, crowd-sourced and user-based innovation has become more important, there is very little empirical analysis of the impact of public sector investments in basic knowledge on innovation through this particular channel. While existing work has generally found that the public sector could significantly boost private sector innovation, this work suggests that caution must be exercised in extrapolating these results to settings where innovators are driven by non-pecuniary motivations, such an online peer production. In such settings, this work shows that public sector involvement in basic knowledge might crowd-out rather than encourage community development and knowledge production.