Knowledge Flows Across Firm Boundaries:

Strategic Implications of Openness

A dissertation presented

by

Do Yoon Kim

to

The Strategy Unit at the Harvard Business School

in partial fulfillment of the requirements

for the degree of

Doctor of Business Administration

in the subject of

Strategy

Harvard University

Cambridge, Massachusetts

May 2019

Dissertation Advisor:
Professor Shane Greenstein

Author:
Do Yoon Kim

Knowledge Flows Across Firm Boundaries:

Strategic Implications of Openness

Abstract

This dissertation examines the strategic implications of opening firm boundaries. The first chapter examines how opening country borders and allowing firms to hire migrant inventors can foster innovations that are different from those of local inventors. The H-1B visa cap increase between 1999-2003 provides an exogenous increase in the ability of firms to hire migrants. Firms affected by this shock can hire more Chinese/Indian inventors and file more herbal patents. Such knowledge is subsequently recombined by local inventors. The second chapter examines how opening a company's software intellectual property affects product market performance. I identify causal effects using community driven reverse-engineering events that exogenously open sourced parts of a company's software. I find that open sourcing corporate software can lead to complement innovations that benefit product market sales. Furthermore, I find strategic implications of considering customer heterogeneity. Complement innovations attract a subset of technologically "savvy" users who inform and influence others' purchase decisions. Thus, opening intellectual property may be more valuable in markets with greater levels of information imperfections. The third chapter examines innovations and collaborations between firms and other contributors in the Linux kernel, a large open source software project. I compare the source code structure before and after the emergence of the Android operating system and document changes in contributions and follow-on innovation. I find evidence consistent with a crowding out of non-corporate efforts from increased corporate contributions, specifically for general purpose files. Additionally, I find that corporate created files lead to less follow-on innovation and have more self-contributions.

Table of Contents

Acknowledgements

To say that no dissertation is written alone would be a vast understatement. Throughout the doctoral journey, I am incredibly grateful to have met and worked alongside so many wonderful people.

First and foremost, I would like to thank my committee members: Shane Greenstein, Dennis Yao, Carliss Baldwin, and Prithwiraj (Raj) Choudhury. Thank you, Shane, for setting the foundations for research. Your three questions (What is the research question, why is it interesting/important, how will you identify it) gave me structure to the entire doctoral process. Without them I would have been defenseless against the vast disarray of exploration/exploitation that is research (I only wish I had unlocked these secrets earlier!). Dennis and Carliss have both been tremendous influences, shaping every aspect of how I think about the world. Not only are they my academic role models, but I aspire to be as kind and understanding people as they are. Last but never least, I thank Raj for pulling me aside one seminar to talk about research. Everything I have learned about the publication process, being careful, thorough, and linking empirics to theory I have learned from him[1].

Much of this work was motivated and driven by friends and experts I met through the Free Software Foundation. I thank Molly de Blanc, Matt Lavallee, Don Robertson, John Sullivan, and Theodore Teah for showing me around FSF and listening to my rambling ideas. Thank you, Bradley Kuhn and Deb Nicholson for your interest in this research, connecting me to more knowledgeable people, and more broadly, their work supporting free software. I am also extremely fortunate to have talked with Jim Gettys, Kevin Hayes, Tao Jin, Andy Oram, Aman Singla, James Vasile, and

[1] All errors are not a reflection of my committee members, but are solely on my part.

several anonymous sources. Much of this would not have been possible without the depth and richness of their insights, as well as their understanding of the technical details.

I have also benefited from my HBS doctoral buddies. My cohort, Yo-Jud Cheng and Mike Teodorescu always had my back (one team one dream!). I learned (almost) everything I know about innovation, and econometrics from Yosub Jung and Ohchan Kwon; I don't know how I would have managed without you two. The senior strategy students were always T+1 steps ahead of me and helped me get everything from research to my general life in shape; thank you, Jasmina Chauvin, Cheng Gao, Jin Hyung Kim, Christopher Poliquin, Haris Tabakovic, and Sarah Wolfolds. I also thank my music buddies Thummim Cho, Max Eber, Jan Iyer, Yizhou Jin, Pavel Kiryeyev, Gea Hyun Shin, Shosh Vasserman, Roxy Wang, and Sungik Yang for keeping me sane in rough times. Finally, thank you Moonsoo Choi, Michael Christensen, Alicia DeSantola, Tommy Pan Fang, Alexandra Feldberg, Grace Gu, Hyunjin Kim, Michael Y. Lee, Jihwon Park, Sean Sa-Pyung Shin, and Michael Uy for your helpful feedback.

This section would not be complete without thanking the most important people in my life. My parents Hyo-Jong Kim and Yeong-Shil Joo have always been huge influences throughout this journey. I still remember sitting down with my parents as a kid, me with my homework and my parents with papers they were revising/reviewing. I try to follow in their examples, albeit slower. My sister Ji Yoon Kim helped me feel less alone in graduate school, and in this world (and reminded me in advance of all-important family deadlines!!). Finally, my sincerest thanks to Sujin Elisa Han for your kindness, help, encouragement, and wisdom. Thank you for giving me the tools of historical context and the importance of reading source material, and thank you for all the countless hours I spent talking to you about routers. None of this would have been possible without you.

Introduction

The flow of knowledge is a fundamental driver of innovation and of economic development. Knowledge flows across a firm's boundaries, in turn, are important for firm performance and innovation. This dissertation examines the strategic implications of opening, allowing for knowledge flows across firm boundaries. The first two chapters of this dissertation examine an "opening" of firm borders, both in terms of human capital (Chapter 1), and intellectual property (Chapter 2). The third chapter takes open boundaries as given and examines how the process of innovation differs in that context.

In Chapter 1, I study how the flow of migrant inventors affects innovation. Ethnic migrant inventors may differ from locals in terms of the knowledge they bring to host firms. We study the role of first-generation ethnic migrant inventors in cross-border transfer of knowledge previously locked within the cultural context of their home regions. Using a unique dataset of Chinese and Indian herbal patents filed in the United States, we find that an increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of herbal knowledge at U.S. assignees by 4.5 percent. Our identification comes from an exogenous shock to the quota of H1B visas and from a list of entities exempted from the shock. We also find that ethnic migrant inventors are more likely to engage in reuse of their prior knowledge, whereas knowledge recombination is more likely to be pursued by teams comprising inventors from other ethnic backgrounds.

In Chapter 2, I study the performance implications of opening-up intellectual property. Increasingly, a firm's ability to create value depends on its ability to achieve alignment with its ecosystem of suppliers and complementors. In such contexts, a firm's strategy may encompass "opening" some parts of their intellectual property (IP) to facilitate value creation, but potentially at the cost of decreased value capture. In this paper, I study the role of customer heterogeneity as a determinant of firms' value capture. I hypothesize that openness creates complement innovations which creates value for customers. These complements are valued by a subset of customers who importantly can mitigate information imperfections in the product market. I further hypothesize such strategies will be more effective for those products that face greater information imperfections. I test these hypotheses on a unique dataset of wireless routers characteristics, complement innovation (custom firmware for wireless routers) compatibility, and product reviews. I utilize an exogenous shock to complement innovation compatibility from several exogenous "reverse engineering" events. I find that the availability of complement innovations increases review ratings by 0.67 stars. I find a strong sorting effect: customers who are more likely to use the custom firmware leave more positive reviews. Importantly, there is a strong information effect as well: these users provide more helpful reviews. Consistent with the information imperfections framework, I find that such effects are stronger for enterprise products which face greater uncertainty.

In Chapter 3, I examine the innovation process within an open regime. A significant part of the digital economy is built upon "digital public goods". Despite significant advances in our understanding of the factors that affect corporate provision of digital public goods, the converse question of how corporate participation affects innovation in digital public goods is less understood. In this paper, we test how increased corporate participation affects source code contribution and follow-on innovation in a large open source project. We test whether corporate contributions increase non-corporate contributions or whether they "crowd out" non-corporate contributions. We utilize a unique data set of contribution behavior from a large open source project to test our hypotheses. We find that while corporate contributions attract more contributions (both from other corporations and non-corporate entities), they also lead to less follow-on innovation. Using a shock that increased corporate commercialization incentives, we find that increased corporate contributions have a "crowding out" effect on individual contributions, specifically to corporate

authored general-purpose code. These results suggest that increased corporate participation may limit innovations in online digital contexts.

1. The Ethnic Migrant Inventor Effect: Codification and Recombination of Knowledge Across Borders

Prithwiraj Choudhury, Do Yoon Kim

Abstract

Ethnic migrant inventors may differ from locals in terms of the knowledge they bring to host firms. We study the role of first-generation ethnic migrant inventors in cross-border transfer of knowledge previously locked within the cultural context of their home regions. Using a unique dataset of Chinese and Indian herbal patents filed in the United States, we find that an increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of herbal knowledge at U.S. assignees by 4.5 percent. Our identification comes from an exogenous shock to the quota of H1B visas and from a list of entities exempted from the shock. We also find that ethnic migrant inventors are more likely to engage in reuse of their prior knowledge, whereas knowledge recombination is more likely to be pursued by teams comprising inventors from other ethnic backgrounds.

Keywords: skilled migration; ethnic migration; first-generation migrant; cultural context; knowledge flows; knowledge reuse; knowledge recombination; recombinant creation; H1B visas

1.1 Introduction

We live in an age of global mass migration on the part of skilled knowledge workers. Recent literature on skilled migration has pointed out that the number of U.S. migrants with tertiary degrees rose by nearly 130 percent between 1990 and 2010 (Kerr *et al.*, 2016). However, this phenomenon coexists with significant concern about migrants displacing jobs for the native-born, and about whether countries should reduce their intake of skilled migrants. This paper sidesteps the debate on whether skilled migrants create or pre-empt employment opportunities for locals. Instead, we study an interesting phenomenon related to the role of ethnic migrant inventors in innovation at their host firms. We argue that ethnic migrant inventors differ from locals and play an important role in transferring to the host firm, knowledge previously *locked within* the cultural context of their home regions. We study recombination of such knowledge once it has been transferred to the host firm.

To motivate this line of research, we present a stylized example involving the role of Dr. Hari P. Cohly, an ethnic skilled migrant and a researcher in immunology at the University of Mississippi, in transferring knowledge about Indian herbal medicine to the west. Dr. Cohly migrated from India to enroll at the University of Toronto and later studied and worked at SUNY Buffalo and the Johnson Space Center in Houston. At the University of Mississippi, Dr. Cohly met Dr. S.K. Das, a plastic surgeon who was about to amputate the leg of a patient with a wound that would not heal due to a condition known as restenosis, a gap between two blood vessels. In his native city of Agra, Dr. Cohly had attended the Indian herbal medicinal (*Ayurveda*) discourses of Dr. M.B. Lal Sahab, a religious teacher and Edinburgh-educated parasitologist. Drawing on these discourses, Dr. Cohly suggested using turmeric to heal the wound. When the patient recovered, avoiding amputation, Dr. Cohly and Dr. Das conducted a clinical trial and filed a U.S. patent for "a wound-healing agent consisting of an effective amount of turmeric powder." This example illustrates the role of an ethnic migrant researcher in codifying knowledge of Indian medicinal herbs and transferring it to the west.

Economic history is replete with such examples of skilled ethnic migrants' transfers of knowledge previously locked within the cultural context of their home regions to the host region and to their host organizations. Scholars have documented the case of Russian mathematicians' transfers of knowledge about such fields as partial differential equations and symplectic topology to the United States, after the fall of the Soviet Union in the 1990s (Borjas and Doran, 2012; Ganguli, 2015). Another example is the seventeenth-century migration of Huguenots from France to Brandenburg-Prussia. As Hornung (2014) states, the Huguenots introduced a great variety of advanced skills and new technologies to their host region. One account lists 46 professions introduced by Huguenots, mostly textile-related, all of which were previously unknown to the host country. One Huguenot carried with him the secret of dyeing fabrics in a special way; another brought the art of printing on cotton. Economic history has also documented the role of the Venetian city council in luring skilled migrant artisans from Florence by offering them patents, which protected their monopoly rights to the knowledge they possessed (Belfanti 2004). After the expiration of the patent, the Florentine artisans joined the local artisan guild in Venice and shared their knowledge with local artisans.

Despite these and other examples (discussed in a later section) of skilled ethnic migrants transferring to the host region knowledge previously locked within the cultural context of their home regions, the recent literature on high-skilled immigration has focused on the employment effects of skilled immigration, namely whether skilled migrants create or displace jobs for the native-born (Kerr, Kerr, and Lincoln, 2015; Kerr and Lincoln, 2010; Doran, Gelber, and Isen, 2016). This question is part of a larger debate in the strategy-and-innovation literature on the global sourcing of talent. Scholars such as Lewin, Massini, and Peeters (2009) have long argued that the shortage of highly skilled technical talent in the United States, and the need to access qualified knowledge workers abroad, should drive western firms' offshoring innovation decisions. This literature tends to take for granted that migrant inventors *resemble* (and thus can replace) local inventors; the literature thus overlooks the possibility that migrant inventors *differ from* local inventors. Specifically, migrant inventors may differ from local inventors in terms of the knowledge they transfer across borders—knowledge that could subsequently be recombined by other inventors, including local inventors.

The field of innovation and strategy has produced a rich literature on the role of context in shaping innovative outcomes (Hambrick and MacMillan, 1985). We draw on this literature, and the literature on cross-national variation in context along cultural, linguistic, religious, and other dimensions (Ghemawat, 2001; Berry, Guillén, and Zhou, 2010), to posit that knowledge can be deeply embedded in its cultural, linguistic, or religious context. We also draw on the literature on impediments to the transfer of sticky knowledge (Von Hippel, 1994; Szulanski, 1996) to argue that, ex-ante, such knowledge can be locked in its home region and thus unavailable to knowledge production in an ethnic migrant inventor's host country.

Given this argument, we study the role of ethnic migrant inventors, especially first-generation ethnic migrants, in transferring such knowledge from their home regions to their host firms. We use the term *home regions* to denote the regions where the knowledge in question was first discovered and/or put into use. We use the phrase *ethnic migrants* solely to denote residents of those regions who migrate to the host firm. Such individuals could be first-generation ethnic migrants (i.e., individuals who migrate themselves) or later-generation ethnic migrants (whose forebears migrated). For purposes of this paper, the phrase *non-ethnic inventors* refers to individuals who did not migrate to the host firm from the home regions of the ethnic migrant inventors. They include both local inventors and those who migrated to the host firm from other regions of the world.

We argue that first-generation ethnic migrant inventors have both the absorptive capacity (Cohen and Levinthal, 1990) and the career incentives (Holmström, 1999) to transfer to their host firms, knowledge that was previously locked within the cultural contexts of their home regions. We also draw on literature on the mechanisms of knowledge recombination (Allen 1977, Fleming 2001) to study knowledge recombination once such knowledge has been transferred to the host firm. We argue that ethnic migrants, whose career incentives stress quick wins (Amabile and Kramer, 2011), tend to reuse prior knowledge, while teams whose members include non-ethnic inventors respond to incentives for knowledge recombination.

To test these propositions, we created a unique dataset of 758 granted herbal patents filed with the U.S. Patent Office (USPTO) between 1977 and 2013 by U.S. firms and universities that were operating during a 1999–2003 immigration shock. The filing entities included such large western multinationals as Abbott Laboratories, Amgen, Eli Lilly, Pfizer, Colgate Palmolive, Dow Chemical, Proctor & Gamble, and Unilever, as well as large U.S. universities. Access to high skilled migrants is important in this industry, as evidenced by responses to the tightening of U.S. immigration policies: for instance, Kenneth Frazier, the CEO of Merck, asserted that "We have to get the best scientists, the best employees around the world" (Johnson, 2017).

Our identification strategy is driven by an exogenous shock to U.S. H1B employment visas. In 1998 and 2000, Congress passed legislation which temporarily increased the quota of H1B visas. The new legislation also exempted universities and a select list of other entities from the quota. Exploiting this policy change to estimate a difference-in-differences model, we find that an increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of herbal knowledge at U.S. assignees. We observe a 4.5 percent increase in herbal patenting at firms subject to the visa cap.[2] Furthermore, inventors of Chinese and Indian ethnicities are more likely to engage in knowledge reuse; meanwhile, teams comprising non-ethnic inventors are more apt to engage in knowledge recombination (i.e., combining herbs with other synthetic compounds to create relatively novel formulations). To our knowledge, we are the first scholars to exploit the H1B exclusion list for purposes of research.

Our findings contribute to the literatures on skilled migration and ethnic migration (Kerr, 2008; Foley and Kerr, 2013; Franzoni, Scellato, and Stephan, 2014) by suggesting that ethnic migrant inventors differ from local inventors with regard to the knowledge they bring to a firm, knowledge that can subsequently be recombined. This insight should be incorporated into policy debates on skilled migration. Our findings also inform the strategy-and-innovation literature on inventor mobility and knowledge flows (Breschi and Lissoni, 2009; Oettl and Agrawal, 2008; Rosenkopf and Almeida, 2003; Song, Almeida, and Wu, 2003) and on the microfoundations of knowledge recombination within firms (Carnabuci and Operti, 2013; Fleming, 2001; Gruber, Harhoff, and Hoisl, 2013).

1.2 Theory and Hypotheses

Ethnic migrant inventors and transfer of knowledge across borders

The strategy and international business literature have documented cross-national variation in contexts along cultural, linguistic, religious and other dimensions (Ghemawat 2001, Berry *et al.* 2010). We build on this literature to argue that knowledge can be *locked* within a given cultural context, and, as a corollary, that such knowledge may not transfer easily to regions whose cultural contexts are dissimilar.

[2] It is plausible that demand side effects lead to an increase in herbal patenting over time; however, given that the visa shock only affects capped firms, our analysis is focused on estimating the effect of a supply shock on herbal patenting.

To make this argument, we draw on the literature on impediments to transferring sticky knowledge (Von Hippel, 1994; Szulanski, 1996; Szulanski, 2002 Jensen and Szulanski, 2004). Szulanski (2002) outlined several impediments to knowledge transfer having to do with the characteristics of the knowledge, including that it might be ex-ante unproven in a new context and/or that causal ambiguity might prevail about the effectiveness of employing it. Knowledge might be considered unproven in certain locales because it is codified in a different language. In the case of Chinese and Indian herbal medicine, western researchers are apt to lack access to knowledge codified in Mandarin and Sanskrit textbooks; they might also be skeptical about using such knowledge, given the paucity of clinical trials proving the effectiveness of herbs. Such knowledge might also suffer from *causal ambiguity*, Szulanski's (2002) term for lack of knowledge about why a given set of actions results in a given outcome ("know-why"). In the case of Chinese and Indian herbal medicine, uncertainty might be concentrated around whether Chinese/Indian herbs would be effective in the western climatic environment. Uncertainty of this kind could prevent effective knowledge transfer. We also leverage the theory of tacit knowledge (Polanyi, 1966, Dasgupta and David, 1994, Cowan and Foray, 1997) to argue that the "know-how" involved in using knowledge embedded in the cultural context of a given country may not transfer easily across borders. Specifically, tacit know-how might be embedded in the actions of expert practitioners and thus available only to immediate observers. Such experts might all be concentrated in the home region, where the knowledge is locked.

Ethnic inventors who migrate to a new region could be instrumental in transferring such knowledge across borders. We argue that ethnic migrant inventors have the absorptive capacity to gainfully employ knowledge previously locked within the cultural context of their home regions. Building on Cohen and Levinthal (1990), we argue that, through prolonged prior exposure, ethnic migrant inventors are likely to have acquired deep understanding of both the know-why and the know-how aspects of knowledge locked in the cultural context of their home regions. They may also have had access to experts who possessed pertinent tacit know-how. In addition to possessing the absorptive capacity to transfer knowledge across borders, ethnic migrant inventors have incentives to codify such knowledge at their host firms. We build on theory about the career incentives of individuals within firms (Holmström, 1999) to argue that transferring and codifying knowledge from their home regions enables ethnic migrants to establish reputations as high performers at their host firms.[3] It is important to note that both the absorptive capacity to transfer knowledge across borders and the incentives to codify such knowledge at the host firm are apt to be especially marked among *first-generation* ethnic migrant inventors. Compared to their second-generation (or longer-settled) ethnic migrant counterparts, first-generation ethnic migrants have had more home-country exposure to pertinent knowledge and to experts who harbor such knowledge and have stronger incentives to distinguish themselves at their new host firms. In other words, when first-generation ethnic migrant inventors move across borders, we can expect to see an increase in the codification of knowledge previously locked within the cultural context of their home countries.[4] This leads us to our first hypothesis:

Hypothesis 1: An increase in the supply of first-generation ethnic migrant inventors increases the rate of codification at their host firms of knowledge previously locked within the cultural context of their home region.

Recombination of knowledge

[3] Holmström (1999) outlines how an individual's concern for a future career may influence his or her incentives to put in effort or make decisions on the job. In the model, the person's productive abilities are revealed over time through observations of performance, and an implicit contract links today's performance to future wages.

[4] In addition to directly codifying such knowledge, first-generation ethnic migrants might transfer it to ethnic and non-ethnic peers at their host firms, who might over time codify variants of it. This insight builds on prior research tracing how mobile inventors might act as cross-cultural brokers, imparting knowledge to co-located peers after geographic mobility (Burt 1992; Oettl and Agrawal, 2008; Almeida and Kogut, 1999; Singh 2005).

Next, we consider how knowledge previously locked within the cultural context of a home region gets *recombined* after its transfer to the host firm, as well as the roles of ethnic migrant inventors and non-ethnic inventors in knowledge reuse and recombination.

The study of knowledge recombination has a rich tradition in the fields of economics and strategy (Schumpeter, 1939; Nelson and Winter, 1982; Henderson and Clark, 1990). One stream of this literature focuses on the microfoundations of knowledge recombination, i.e., the role that individuals play in knowledge recombination within a given firm. This tradition dates back to Allen (1977) and is framed by Fleming (2001) as the process of recombinant search that is characteristic of individual inventors. In the subsequent literature, Carnabuci and Operti (2013) designate two distinct recombinant search strategies: "recombinant creation" (creating recombinations new to the firm) and "recombinant reuse" (reconfiguring combinations already known to the firm). In our context recombination is more likely to consist of recombinant creation, or recombining knowledge components from the western context with knowledge components previously embedded in the ethnic migrant inventor's home region. We build on this literature to theorize that, in the case of knowledge transferred from the cultural context of a home region to a host firm, ethnic migrant inventors are more likely to engage in knowledge reuse than in either form of recombination; by contrast, teams that include non-ethnic inventors are more likely to engage in recombinant creation.

Building on the prior literature about absorptive capacity and career incentives, we argue that the incentives driving ethnic migrant inventors tend to promote reuse of knowledge over recombination. In our setting, we frame knowledge reuse as the ethnic migrant inventor's appropriation of knowledge from the home country and codification of that knowledge in the new context of a western firm. Given her relatively high absorptive capacity with respect to knowledge previously locked in her home country, the ethnic migrant can realistically expect a high likelihood of success at reusing home-country knowledge. We build on expectancy theory (Vroom, 1964) to argue that this expectation is likely to motivate reuse of such knowledge by ethnic migrants. Also, ethnic migrant inventors (especially first-generation ethnic migrants) at host firms experience strong incentives to establish a reputation quickly. The literature on career incentives has shown that *quick wins* help boost individuals' confidence and instill a sense of career progress (Amabile and Kramer, 2011; Connelly *et al.*, 2011). Arguably, transferring and reusing home-country knowledge could be a quick win for ethnic migrants. In contrast, recombining the same knowledge with components well known to the western firm would require the ethnic migrant inventor to develop the absorptive capacity to assimilate knowledge components from *both* settings. We also know from prior literature that developing the absorptive capacity to work with new knowledge components calls for repeated exposure to related problems over the course of many practice trials. Hence, recombination is unlikely to result in a *quick win* for ethnic migrants.

We also theorize that non-ethnic inventors might be more inclined to engage in recombinant creation by working with ethnic migrant inventors and/or by working on their own. As prior literature has shown, to engage in recombinant creation, inventors need to have knowledge diversity (Cohen and Levinthal, 1990; Ahuja and Morris Lampert, 2001; Carnabuci and Operti, 2013).[5] When non-ethnic inventors collaborate with ethnic migrants, the resulting team would have greater knowledge diversity than either of the inventor groups working separately. Recombinant creation could also emerge from non-ethnic inventors working alone. Non-ethnic inventors might learn the new knowledge being codified from their ethnic peers and might recombine such knowledge. This insight builds on prior research tracing how mobile inventors might act as cross-cultural brokers,

[5] According to Cohen and Levinthal (1990), knowledge diversity facilitates the innovative processes of individual inventors by promoting novel associations and linkages pertinent to the problems they are attempting to solve. In a similar vein, Ahuja and Morris Lampert (2001) have shown that knowledge diversity helps individuals find radically novel approaches to solving technological problems. Extending this argument, Carnabuci and Operti (2013) theorize that knowledge diversity helps individual inventors engage in recombinant creation.

imparting knowledge to co-located peers after geographic mobility (Burt 1992; Oettl and Agrawal, 2008; Almeida and Kogut, 1999; Singh 2005). However, working on a recombinant creation project is a risky choice for an individual. As Holmström (1999) states, the risk preferences that govern choices of projects are driven by individuals' career concerns. Non-ethnic inventors might make the risky choice of working on recombinant creation, given that information about their talent is more likely to be known to the firm. In short, we expect higher reuse to emerge from teams composed exclusively of ethnic migrants. By contrast, we expect recombinant creation to emerge from teams whose members include non-ethnic inventors. This leads to our second hypothesis:

Hypothesis 2: Teams composed solely of ethnic migrants are more likely to reuse knowledge from their home regions at their host firms; knowledge recombination is more likely to be pursued by teams that include non-ethnic inventors.

1.3 Data, Variables, and Identification Strategy

To test our hypotheses, we use a unique dataset of Chinese and Indian herbal patents filed in the United States. For several reasons, herbal patents are an appropriate empirical setting in which to study the transfer and recombination of knowledge previously locked within the cultural context of ethnic migrant inventors' home regions. For centuries, both China and India have accumulated extensive knowledge about the medicinal properties of herbs, within medical canons distinct from the western medical canon (*Ayurveda, Unani, Siddha, Yoga*, and TCM, or Traditional Chinese Medicine). Chinese and Indian migrant knowledge workers account for more beneficiaries of temporary U.S. work visas than any other national groups. In the pharmaceutical industry, which generates many herbal patents, immigrants represent 33% of the total R&D workforce and 43% of medical/life scientists (Michel and Witte, 2014). This scenario represents an opportunity to test whether knowledge of Chinese and Indian herbal medicine is transferred to the west by first-generation migrant Chinese and Indian scientists, and whether non-ethnic inventors play an important role in recombination of that knowledge.

A unique dataset of herbal patents

Within the entire universe of USPTO patents, we identified herbal patents filed between 1977 and 2013. We categorized a patent as herbal if the application named at least one herb and specified its use. Our search process consisted of three iterative steps. First, we first obtained a list of 52 herbs, and their common and scientific names, from the National Center for Complementary and Alternative Medicine (NCCAM) website. We then searched Thompson Innovation and LexisNexis TotalPatents for USPTO patents whose abstract or title named any of these herbs. We found 7,163 such patents. We then iteratively searched the identified patents for more herb names and collected additional herbal patents. The addition of herb names extracted from those patents ultimately produced a list of 1,785 herbs. The total number of patent-herb pairs exceeds the number of herbal patents because a single patent can name multiple herbs. The most frequently named herbs appear in Table A17 in the Appendix. Next, we performed a classification search using both the International Patent Classification (IPC) and the U.S. Patent Classification (USPC) schemes.[6] Finally, we used the Traditional Chinese Medicine (TCM) database to augment our dataset and read patent abstracts to further validate our list and identify additional herb names and their uses; we appended all patents with U.S. priority to our existing dataset.

We then collected information about the patent assignees from USPTO's PatentsView and Capital IQ. Our identification strategy, detailed below, focused on U.S. entities that *could have* hired inventors between 1999 and 2003; thus, we concentrated on patents filed by U.S. assignees (firms

[6] In particular, we used the IPC class A61K36+ and the USPC classifications 424/725 and 514/783. The IPC class was introduced in 2002 by a Committee of Experts at IPC Union for purposes of linking the Traditional Knowledge Research Classification (TKRC) with IPC as part of the work of the World Intellectual Property Organization Traditional Knowledge (WIPO-TK) Task Force.

and universities) that could have hired inventors in 1999–2003. To determine which assignees fell into this category, we used PatentsView's assignee classification data and data from Capital IQ on the locations of the assignees' headquarters. Our sample consists of 1,794 patent filings filed by U.S.-based assignees, 981 of which were granted. Capital IQ also provided firms' founding dates, allowing us to impute a firm's age.[7] We dropped 536 patents whose assignees (mostly firms) were founded after the visa shock (and thus not affected by it) or that stopped patenting before 1999, the first year of the shock, suggesting that the assignee may not have been in operation during the visa-shock period. Doing so left 1,258 patent filings.[8] In the base case, we kept only patent applications that were granted, resulting in 758 patents submitted by 401 assignees.

Finally, we aggregated our herbal patent data at the assignee-year level by counting the number of herbal patents granted to a given assignee in each year. We used the *tsfill* command in Stata to fill in missing assignee-year-level observations. For years when an assignee was not granted any herbal patents, we set the assignee-year observation at zero.[9] We dropped any assignee-year pairs that corresponded to dates before an assignee's founding date. The assignee-year level dataset is thus an unbalanced panel consisting of 8,998 observations (an average of 22.4 years of observations for 401 assignees).

[7] For assignees without an equivalent firm in Capital IQ, we used the earliest patent application year in the USPTO database as the assignee's founding year. Our results are robust to other definitions of assignee founding years, and to including the assignee in the sample, even prior to the first patent filing date, with lags.

[8] This number includes patents filed by entities that did not file patents during the visa shock but did so after it ended. Our results are robust to using the sample of assignees that filed patents during the visa-shock period. Section 7.5 in the appendix contains more details on the selection of assignees.

[9] The outcome variable, number of herbal patents granted to an assignee in a given year, has many zeros; and the variance of this variable is larger than its mean. Consequently, we investigated whether a Poisson regression is appropriate for this data. Following Cameron and Trivedi (2010), we checked whether the outcome variable looks like a Poisson-distributed random variable. We see a slightly higher probability of zeros in the observed outcome than the Poisson distribution would predict (0.0089), but no difference in predicted counts of zeros using Stata's countfit command. The contribution of zeros to the Pearson Chi-Square statistic is 0.001, further showing that our data does not suffer from over inflation of zeros. We also ran robustness checks using a Poisson and Poisson QML specification.

1.3.1 Identification strategy

We proposed two hypotheses: that an exogenous increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of knowledge previously locked within their home countries; and that while ethnic migrant inventors are more likely to reuse knowledge transferred across borders, recombination is more likely to be pursued by teams comprising non-ethnic inventors. To test the first hypothesis, we run a difference-in-differences model on log herbal patent counts at the assignee level to determine whether herbal patenting changes as the supply of Chinese/Indian migrants to the U.S. changes. To test the second hypothesis, we test whether the probability of recombination correlates with the presence of Chinese/Indian inventors' names on a patent. The next section describes our natural experiment, defines variables, and presents empirical specifications.

A Natural Experiment: The H1B Visa Shock and Excluded Entities

The key barrier to identifying whether or not a supply shock of first-generation ethnic migrants leads to greater codification of knowledge is the existence of unobservables affecting both codification and immigration patterns. In our setting, the returns to investment in herbal patenting increased in the mid-1990s as the market for herbal remedies grew. This increase in demand may have led firms to accelerate herbal patenting and to recruit more experts on herbal remedies. Also, even if we found a correlation between ethnic migrant inventors and codification of knowledge, we would be unable to determine whether these ethnic inventors were first-generation migrants. The goal was to find an exogenous increase in the inflow of ethnic Indian and Chinese inventors to the United States, unrelated to determinants of herbal patenting.

In pursuit of this goal, we utilized an exogenous shock to skilled immigration to the United States. In 1998 and 2000, Congress promulgated two laws that differentially impacted some firms' capacity to recruit skilled labor from abroad. As a result, the number of H1B employment visas increased from 65,000 in 1998 to 115,000 in 1999 and 195,000 in 2001, and then dropped back to 65,000 in 2004. Both laws were responses to increased demand for information technology (IT) professionals during the dot-com bubble. Thus, the flow of first-generation migrants is plausibly exogenous to the filing of herbal patents since most of the workers filled IT-related positions. We focus on Chinese and Indian inventors because they are the two largest groups to receive H1B visas: workers from India account for the majority of H1B recipients, followed by workers from China. Figure 1.1 illustrates the cap on H1B visas over time. In the appendix (section 2.1), we verify that this H1B shock was meaningful for our sample of assignees by comparing the trend in Labor Condition Applications (LCAs)—a prerequisite filed by an employer that intends to apply for H1B visas—submitted by our sample of assignees to the number of *new* unique ethnic names on the list of inventors in our sample during the shock period.

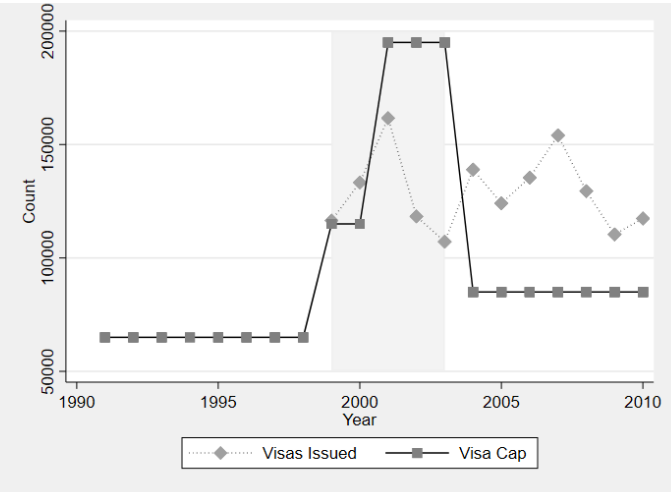


Figure 1.1. H1B visa cap over time

Notes: Figure 1.1 plots the H1B visa cap and visa issuances over time. The shaded area represents the period during which the American Competitiveness in the 21st Century Act (AC21) was in effect. In accordance with the AC21 Act, H1B visa quotas were raised between 1999 and 2003 and lowered starting in 2004. The American Competitiveness and Workforce Improvement Act (ACWIA) passed in 1998 increased the H1B visa cap from 65,000 to 115,000. The American Competitiveness in the 21st Century Act (AC21) passed in 2000 further increased the visa cap to 195,000. AC21 also retroactively increased the 1999 and 2000 quotas above the 115,000 cap set by the ACWIA. The actual H1B visa issuances in 1999-2003 were 116,513 (1999), 133,290 (2000), 161,643 (2001), 118,352 (2002), and 107,196 (2003), all greater than the pre-1999 cap of 65,000. Visa issuances post-2003 are above the visa cap, as AC21 created the cap-exempt group.

An interesting feature of this immigration shock, to our knowledge not previously exploited in academic research, is that certain entities were exempted from the visa cap: workers "(1) at an institution of higher education or a related or affiliated nonprofit entity, or (2) at a nonprofit research organization or a governmental research organization"[10] could hire as many employees as they wished via the H1B visa. In addition to universities, the cap-exempt assignee list includes firms like Amgen, Monsanto, and Pioneer.[11] Of our 401 assignees, 73 were exempt from the visa cap; they were granted 123 patents. This exemption allows us to study the differential effect of the visa-cap increase by comparing herbal patent grants at capped and exempt patenting assignees.

1.3.2 Dependent Variables

Patent counts

Our main dependent variable (*Log Patent Count*) is the log number of herbal patents granted to an assignee (a firm or university) in a given year. We aggregate our herbal patents into assignee-year-level observations. For years in which an assignee was included in the sample but was not granted any herbal patents, we set the number of patents at zero. Since the number of herbal patent grants in an assignee-year is skewed, we add 1 and take logs.[12]

Recombined

To create the dependent variable for Hypothesis 2, we code herbal patents as *Recombined* if the patent text refers to synthetic non-herbal formulations in addition to herbs. We use the Derwent

[10] Source: https://www.uscis.gov/sites/default/files/USCIS/Laws/Memoranda/Static_Files_Memoranda/Archives%201998-2008/2006/ac21c060606.pdf

[11] Firms were able to hire migrants in the 'cap-exempt' mode because the statute exempted from the numerical limitations of the cap those migrants who were employed "at" a qualifying institution, a broader category than employment "by" a qualifying institution. In other words, firms could also claim cap exemption in hiring a migrant because the migrant would perform duties "at" a qualifying institution.

[12] Our results are robust to using nonlinear count models, including negative binomial, Poisson and Poisson QML. As discussed earlier, our data does not suffer from over inflation of zeros.

classification, a manually curated standardized system for classifying patents maintained by Thomson Reuters. We identify three classes within the Derwent classification that consist of both herbs and synthetic compounds, and code a patent as recombined if it belongs to any of these classes. Section 7.1 in the Appendix explains how this variable was coded.

New inventors

PatentsView allows us to identify individual inventors in our dataset and to track them across time. We combine this data with our ethnicity classification, described below, to count the number of new Chinese/Indian inventors associated with each patent (*New Ethnic Inventors*).

1.3.3 Independent Variables

Inventor ethnicity

Though patent documents do not specify inventors' ethnicities, we were able to predict likely ethnicities using names' linguistic cues. Probabilistically, surnames like Xing are more likely to be associated with Chinese individuals than with other ethnicities. Building on this insight, we utilized the open-source name categorizer "ethnicityguesser" to determine inventors' ethnicities.[13] The software, based on the Natural Language ToolKit (NLTK) package in Python, comes prepackaged with a set of names and associated ethnicities. As a robustness check, we compared our ethnicity-classification results when using different training sets to Ambekar *et al.* (2009), who use state-of-the-art hidden Markov models and decision trees for classification. The Appendix reports correlations across our measures and other established measures of ethnicity classification (see Tables A12 and A13). We find correlations above 0.9 for all our ethnicity measures. We create indicator variables to

[13] GitHub kitofans/ethnicityguesser - https://github.com/kitofans/ethnicityguesser

specify whether the set of inventors on a patent is uniformly Chinese/Indian (*Fully Ethnic*), uniformly non-Chinese/Indian (*Non-Ethnic*), or of mixed ethnicity (*Mixed Team*).

1.4 Empirical Specifications

To recap, Hypothesis 1 states that an increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of knowledge previously locked within the cultural context of migrant inventors' home regions. Our identification comes from a difference-in-differences model using capped firms as the treated group; the treatment period is the visa-shock period. We estimate this regression equation using our assignee-year level data:

$$\boldsymbol{ln\left(1 + Patent\ Count_{jt}\right) = \alpha + \beta_1 Capped_j + \beta_2 Shock_t + \gamma Capped_j \times Shock_t + \phi_j + \lambda_t + \varepsilon_{jt}} \qquad (1)$$

Our dependent variable $ln(1+Patent\ Count_{jt})$ is the log number of herbal patents filed by assignee *j* in year *t*. The variables $Capped_j$, $Shock_t$ are dummies for whether assignee *j* is subject to the H1B cap and whether the patent application year is within the treatment period, 1999–2003. To control for assignee-level unobservables and temporal unobservables, we include assignee fixed effects ($\boldsymbol{\phi_j}$) and year fixed effects (λ_t) in some specifications, in which case the variables $Capped_j$, $Shock_t$ are respectively dropped. Here, the β coefficients capture the time-invariant difference in patenting between capped firms and exempt firms (β_1), and the percentage change in herbal patenting over time (β_2). Our main coefficient of interest, the interaction term γ, captures the percent increase in herbal patenting caused by relaxing the H1B visa cap.

To test Hypothesis 2, which asserts that teams composed solely of ethnic migrants are more likely to reuse knowledge than to recombine it, we must document the relationship between knowledge recombination and inventor ethnicities. We test whether there is a significant association between a patent with only Chinese/Indian inventors and the nature of the knowledge created (i.e., recombination or reuse). We run the following regression equation using our patent-level data:[14]

$$\boldsymbol{Recombined_{ijt} = \beta Fully\ Ethnic_{ijt} + \phi_j + \lambda_t + \varepsilon_{ijt}} \quad (2)$$

Again, the dependent variable (*Recombined$_{ijt}$*) is coded as 1 if patent *i* by assignee *j* with application year *t* contains synthetic herbal compounds. *Fully Ethnic$_{ijt}$* signifies that a patent's inventors all have Chinese/Indian names. We include assignee fixed effects (ϕ_j) to control for assignee-level unobservables that determine the likelihood that a patent mentions synthetic compounds. The *β* coefficient measures the association between ethnic composition and the probability of knowledge recombination. Because Hypothesis 2 states that patents with exclusively ethnic inventors are more likely to involve reuse, we should see *β* < 0 even after controlling for assignee fixed effects.

We are also interested in temporal trends in recombination. Specifically, we wish to learn whether, over time, knowledge is more likely to be recombined, and by which ethnic groups. In the base case, we estimate the following equation:

[14] We also utilize the supply shock and estimate alternate specifications: $\text{Recombined}_{\text{ijt}}=\delta\text{Capped}_{\text{j}}\times \text{Shock}_{\text{t}} +\lambda_{\text{t}} +\phi_{\text{j}} +\varepsilon_{\text{ijt}}$. and $Recombined_{ijt} = \gamma Capped_j \times Shock_t \times FullyEthnic_i + \beta_1 Capped_j \times Shock_t + \beta_2 Shock_t \times FullyEthnic_i + \beta_3 Capped_j \times FullyEthnic_i + \alpha FullyEthnic_i + \lambda_t + \phi_j + \varepsilon_{\text{ijt}}$. We include year fixed effects (λ_t) and assignee fixed effects (ϕ_i). δ measures the change in likelihood of recombination caused by an increase in the flow of first-generation ethnic migrants, and γ measures the differential impact of the shock on patents with solely ethnic inventors.

$$\boldsymbol{Recombined_{iht} = \alpha + \beta_1 Time\ Since\ Herb\ Introduced_{iht} + \beta_2 (Time\ Since\ Herb\ Introduced_{iht})^2 + \nu_h + \varepsilon_{iht}} \qquad (3)$$

Our dependent variable is an indicator denoting whether patent i filed at time t uses herb h, and whether it is a recombination of herbs and synthetic compounds. Our independent variable measures how much time has passed since herb h was first introduced in the United States, measured using the time between the filing date of the first U.S. patent containing the herb and the filing date of patent *i*. We include herb fixed effects ν_h because we are interested in the within-herb effect of time. The β_1 coefficient with herb fixed effects thus measures, for a particular herb, whether it is more likely to be recombined initially ($\beta_1 < 0$) or after it has become more widespread ($\beta_1 > 0$). The β_2 coefficient estimates non-linear effects.

1.5 Results

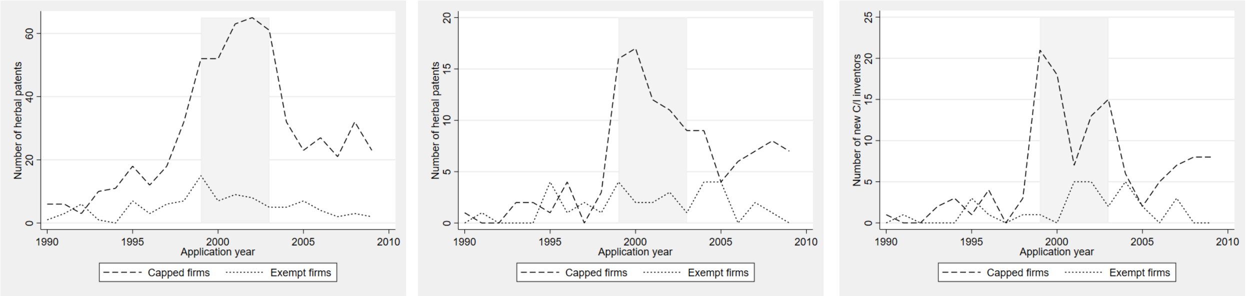
Summary statistics for capped and exempt assignees

Table 1.1 compares the assignee-year-level characteristics of capped and exempt assignees, during and outside of periods characterized by a visa-cap increase. We report the means and standard deviations, and the results of a *t*-test with standard errors. Figures 1.2a and 1.2b document the increase in herbal patenting by U.S. entities during our study period: we see a significantly greater increase in herbal patenting between 1999 and 2003, when the visa cap was increased, at capped firms than at exempt firms. In Figure 1.2c, we observe a similar trend in the numbers of new Chinese/Indian inventors' names on patents. We will verify these trends below using robust econometric methods.

Table 1.1. Summary statistics

Capped	(1) Normal visa cap		(2) Increased visa cap		(3) Diff (2) – (1)	
	mean	sd	mean	sd	b	Se
Patent counts	0.06	0.29	0.21	0.55	0.15	(0.02)
Recombined	0.02	0.14	0.07	0.34	0.05	(0.01)
New inventors	0.11	0.66	0.39	1.42	0.28	(0.04)
New ethnic inventors	0.01	0.14	0.05	0.34	0.04	(0.01)
Ethnic inventors	0.35	0.85	0.38	0.83	0.04	(0.07)
Observations	5470		1384		6854	
Exempt	(4) Normal visa cap		(5) Increased visa cap		(6) Diff (5) – (4)	
	mean	sd	mean	sd	b	Se
Patent counts	0.04	0.21	0.12	0.41	0.08	(0.02)
Recombined	0.01	0.10	0.05	0.28	0.04	(0.02)
New inventors	0.10	0.58	0.24	1.00	0.14	(0.05)
New ethnic inventors	0.02	0.22	0.04	0.29	0.02	(0.02)
Ethnic inventors	0.57	1.07	0.56	1.08	-0.01	(0.22)
Observations	1788		356		2144	
Difference-in-Differences					Diff-in-Diff (6) – (3)	
					b	Se
Patent counts					0.07	(0.02)
Recombined					0.01	(0.01)
New inventors					0.14	(0.05)
New ethnic inventors					0.02	(0.01)
Ethnic inventors					0.05	(0.20)
Observations					8998	

Note: Standard errors appear in parentheses. Observations are at the assignee-year level, for all years that an assignee (firm or university) was in operation. We use the *tsfill* command in Stata to fill in missing assignee-year pairs. The variable "Patent counts" refers to the number of herbal patents filed by an assignee in a given year. The variable "Recombined" measures the number of recombined patents at the assignee-year level. In regressions, we use "Recombined" as an indicator for whether the patent contains synthetic compounds as well as herbal ingredients. The variables "New inventors" and "New ethnic inventors" represent the number of inventors in a firm year who file their first patent, and the same number for inventors with Chinese/Indian names. The variable "Herb count" counts the average number of herbs on a patent.



(a) All herbal patents

(b) Herbal patents by Chinese/Indians

(c) New Chinese/Indian inventors

Figure 1.2. Number of herbal patent grants and new inventors over time
Notes: Shaded areas represent the period during which the visa cap was increased due to the AC21 Act. We see a sharper increase in the number of herbal patents during the visa-shock period, and especially in herbal patents with ethnic inventors for capped firms, compared to exempt firms. The number of new Chinese/Indian inventors also rises significantly during the same visa-shock period, especially for capped firms.

Testing Hypothesis 1: Difference-in-differences estimation

Hypothesis 1 states that an increase in the supply of first-generation ethnic migrant inventors increases the rate of codification of knowledge previously locked within the cultural context of those inventors' home regions. Table 1.2 presents the results of estimating our main difference-in-differences specification, equation (1) using OLS. Our dependent variable is the log of 1 plus the number of patent grants at the assignee-year level. Standard errors are clustered at the assignee level.

Table 1.2. The effect of a visa shock on herbal patents

	(1)	(2)	(3)	(4)
	Dependent variable: log(1+patent count)			
Capped x Shock	0.044	0.038	0.042	0.044
	(0.017)	(0.017)	(0.017)	(0.018)
Capped	0.010	0.016	0.009	
	(0.005)	(0.005)	(0.005)	
Shock	0.048	0.052		
	(0.014)	(0.014)		
Constant	0.030	0.042	0.000	0.027
	(0.004)	(0.006)	(0.010)	(0.010)
Controls	N	Y	Y	Y
Application Year FE	N	N	Y	Y
Assignee FE	N	N	N	Y
Observations	8998	8998	8998	8998
Adjusted R^2	0.030	0.059	0.068	0.072

Note: Standard errors appear in parentheses, clustered at the assignee level. Observations are at the assignee-year level, for all years that an assignee (a firm or university) was in operation. We use the *tsfill* command in Stata to fill in missing assignee-year pairs. The assignee-year-level dataset is thus an unbalanced panel consisting of 8,998 observations (an average of 22.4 years of observations for 401 assignees). The dependent variable is the log of the number of herbal patents filed by an assignee in a given year. *Capped* is an indicator for whether the assignee is subject to the visa cap; *Shock* is an indicator for years 1999 through 2003. Controls include the fraction of inventors who are Chinese/Indian, assignee age, number of inventors, and the number of Chinese/Indian inventors. Percentage increases are calculated as $100 \cdot (e^{\beta} - 1)$. In the base case, we use an OLS specification. For models with assignee fixed effects, we use an xtreg specification. The dependent variable, number of herbal patents granted to an assignee in a given year, has many zeros; and the variance of this variable is larger than its mean. Consequently, we investigate whether a Poisson regression is appropriate for this data. Following Cameron and Trivedi (2010), we check whether the outcome variable looks like a Poisson-distributed random variable. We see a slightly higher probability of zeros in the observed outcome than the Poisson distribution would predict (0.0089), but no difference in predicted counts of zeros using Stata's countfit command. The contribution of zeros to the Pearson Chi-Square statistic is 0.001, further showing that our data does not suffer from over inflation of zeros. We also ran robustness checks using a Negative Binomial, Poisson and Poisson QML specification. Alternatively, in the Appendix (Table A18) we present a DD specification using Post2004 (tightening immigration policies) as the treatment period, which yields a negative and significant coefficient (β=-0.036, p<0.001).

We see from the coefficient on the interaction term (*Capped x Shock*) in Column 1 that the visa shock increased the log herbal patent count (β = 0.044, p = 0.010). This suggests a 4.5 percent increase in herbal patenting at firms subject to the visa cap. The positive coefficient on the $Capped_j$ variable denotes that capped firms file more patents than exempt firms, (β = 0.010 p = 0.026), invariant to time. Similarly, we see from the positive coefficient on *Shock*$_j$ that the visa-shock is associated with an increase in herbal patenting, (β = 0.048 *p* = 0.001). Column 2 controls for assignee-year-level factors that may affect herbal patenting, such as assignee age and the number of Chinese/Indian inventors employed by the assignee in a given year. The coefficient on our main dependent variable is robust to controlling for year fixed effects (Column 3) and for assignee fixed effects (Column 4). For the last specification, we see that the visa shock caused herbal patenting to increase by 4.5 percent, with β=0.044, *p*=0.013. Our results are robust to relaxing the founding-year assumptions discussed in the Appendix (Table A7) and to using nonlinear count models instead of OLS, which we report in the Appendix (Table A8). Separately, in Appendix table A18 we test whether stricter immigration policies after 2004 decreased herbal patenting and find a negative effect (β= -0.036, p<0.001), i.e. a 3.7 percent decrease in herbal patent filing at capped firms after reduction of the visa cap in 2004.

Testing for parallel trends and dynamics of the visa shock

To learn how the visa shock affected the two groups of assignees each year, and to see whether the common-trends assumption holds, we include interactions between capped-firm dummies and lead and lag terms for the implied visa shock. Graphically, we can plot how being in the capped sample affected herbal patenting by Chinese/Indian inventors over time by plotting the coefficients on all interaction terms, as in Autor (2003). Doing so also allows us to compare the treatment and control groups during the pre-treatment period. In Figure 1.3, we see that herbal patenting increased significantly between 1999 and 2003, when the visa cap was increased. Furthermore, with the exception of 1993 and 1994, all the coefficients on the interactions are statistically insignificant in the pre-shock period, supporting the difference-in-difference assumption of parallel trends.[15]

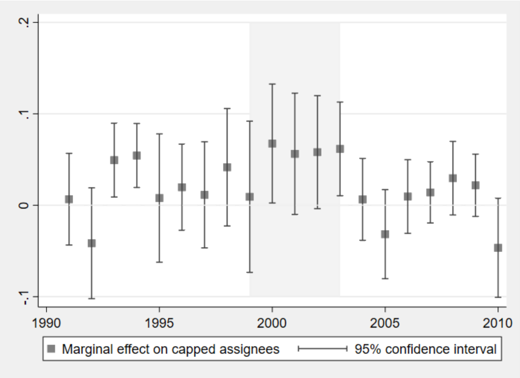


Figure 1.3. Estimated impact of visa shock on herbal patenting for capped versus exempt firms in years before, during, and after 1999-2003 shock

Notes: Standard errors are clustered at the assignee level. The shaded area represents the period during which the visa cap was raised by the AC21 Act. We estimate $ln(1 + Patent\ Count_{jt}) = \lambda_t + Capped_j + \sum_\tau \delta_\tau(\lambda_t \times Capped_j)$ where we include dummies for capped firms ϕ_j, year fixed effects λ_t, and all interactions between year dummies and capped firms. Figure 1.3 plots the coefficients and confidence intervals for all interaction terms δ_τ.

New-herb introduction

We have shown that a supply shock of first-generation migrants increases codification of herbal patents in the United States. Yet we do not know whether the knowledge being codified was unfamiliar to western assignees as suggested by H1, which hypothesized the codification of knowledge previously locked within the home region of ethnic migrant inventors. We estimate an OLS model for the probability that new inventors in the sample, especially new ethnic inventors, introduce a *new* herb and/or an herb unfamiliar in the west (as measured by Google Ngrams). The results appear in Table A4 in the Appendix.[16] We find that new inventors are 12 percentage points more likely than existing inventors to introduce a new herb (*p*= 0.006); the presence of new ethnic

[15] The positive and significant effects of 1993 and 1994 might be related to the Immigration Act of 1990 (enacted on November 29, 1990), which provided for 140,000 visas per year for job-based immigration in five categories (EB1, EB2, EB3, EB4 and EB5). It also created new categories of nonimmigrant visas (the O and P categories) for extraordinarily skilled foreigners in the realm of science.

[16] Specifically, we test $1(HerbIntro)_{ijt} = \beta_0 + \beta_1 NewInventor_i + \beta_2 NewEthnicInventor_i + \phi_j + \lambda_t + \gamma X_{jt} + \varepsilon_{ijt}$, and $NewHerbFamiliarity_i = \beta_0 + \beta_1 NewInventor_i + \beta_2 NewEthnicInventor_i + \phi_j + \lambda_t + \gamma X_{jt} + \varepsilon_{ijt}$

inventors on a patent adds 15 percentage points to the likelihood of introducing a new herb (*p*= 0.041). Similarly, the participation of a new inventor is associated with approximately a one-standard-deviation decrease in the measure of prior familiarity of the herb in the west (*p*= 0.004); the participation of a new ethnic inventor is associated with an additional one-standard-deviation decrease (*p*= 0.047) in prior familiarity. Section 3 in the Appendix provides more detail.

Testing Hypothesis 2: Knowledge reuse and knowledge recombination

Hypothesis 2 states that ethnic migrant inventors are likely to reuse knowledge transferred from their home regions; knowledge recombination is more likely to be pursued by teams that include non-ethnic inventors. Table 1.3 presents the results of estimating Equation 2 using OLS and clustered standard errors, clustered at the assignee level. Column 1 compares patents with only Chinese/Indian names to patents with at least one non-ethnic inventor. We see that fully ethnic teams are 34.6 percentage points less likely to engage in recombination than teams that include non-ethnic inventors (*p*=0.026). Furthermore, we utilize the identification strategy as an exogenous supply shock in the likelihood of observing fully ethnic teams consisting exclusively of first-generation ethnic migrants. We estimate a difference-in-differences specification similar to equation (1) and interact it with an indicator for fully ethnic teams to derive the differential impact of the visa shock on the fully ethnic groups. Column 2 shows how the visa shock affected recombination; column 3 shows the differential impact of the visa shock on patents with only Chinese/Indian names (fully ethnic). Column 4 compares recombination probabilities for mixed teams and non-ethnic teams. All specifications include assignee and application year fixed effects.

Table 1.3. Probability of recombination across ethnicities

	(1)	(2)	(3)	(4)
		Dependent variable: Recombined		
Fully Ethnic	-0.346 (0.154)		-0.609 (0.318)	
Capped x Shock		-0.313 (0.136)	-0.243 (0.132)	
Capped x Shock x Fully Ethnic			-0.897 (0.218)	
Fully Ethnic x Shock			0.872 (0.198)	
Fully Ethnic x Capped			0.293 (0.288)	
Non-ethnic				0.360 (0.155)
Mixed Teams				0.315 (0.162)
Constant	-0.995 (4.688)	0.481 (0.186)	0.425 (0.185)	-1.272 (4.737)
Controls	Y	Y	Y	Y
Application Year FE	Y	Y	Y	Y
Assignee FE	Y	Y	Y	Y
Observations	758	758	758	758
Adjusted R^2	0.062	0.073	0.078	0.062

Note: Standard errors appear in parentheses, clustered at the assignee level. Observations are at the patent level. We retain only herbal patents filed by U.S. assignees in operation during the visa-shock period. The dependent variable is an indicator for whether the patent contains synthetic compounds. *Fully Ethnic* is an indicator for patents with only ethnic inventors. *Capped x Shock* is an indicator denoting whether the patent was filed by a cap-subject firm between 1999 and 2003 (the visa-shock period), and *Capped x Shock x Fully Ethnic* interacts this with an indicator for Full Ethnic teams. *Non-Ethnic* and *Mixed teams* are indicators for patents listing no inventors with Chinese/Indian names and for patents with some but not all Chinese/Indian inventors. The omitted group for column (1) is patents with any non-Chinese/Indian inventors. In column 4, the omitted group is patents by teams composed entirely of Chinese/Indian inventors. Controls include assignee age at the time of application, number of claims listed on the patent, whether the patent had any new inventors, average prior co-inventor centrality, and average prior exposure to ethnic co-inventors. In the base case, we use an OLS specification (given that all models include assignee fixed effects, we use an xtreg specification).

Column 2 shows that inventors at firms that were subject to the visa cap are 31.3 percentage points less likely to recombine knowledge (δ = -0.313, p=0.022).[17] Furthermore, Column 3 shows a larger negative effect for fully ethnic teams at capped firms during the shock (γ=-0.897, p<0.001), suggesting that reuse is being driven by fully ethnic teams consisting of first-generation migrants. Finally, Column 4 shows that both mixed and fully non-ethnic teams are more likely to recombine than are fully ethnic teams (p=0.053 and p=0.021 respectively). A t-test fails to reject the null of no difference between these two coefficients. In each specification, we include various controls. We include the number of claims and the citation count to control for patent characteristics. Since recombination may be driven by the broader collaboration network of prior co-inventions, we control for average number of co-inventors, i.e. average centrality, and for prior exposure to ethnic co-inventors computed at the patent level (results available with authors upon request).

Recombination over time

Figure 1.4 (a) plots the likelihood of recombination over time after an herb is introduced to the United States (measured by the filing date of the first U.S. patent containing the herb), by inventor ethnicity, averaged across all herbs mentioned in the patent text.[18] The solid line plots the probability of recombination by any ethnicity, and is the sum of the dashed lines, which plot the probabilities of recombination by specific ethnicities. We see a slightly positive, nonlinear relationship for recombination in general, especially for non-ethnic inventors. Mixed teams seem to increase their rate of recombination later in an herb's life.

[17] Our results are robust to nonlinear specifications using logistic regression. Results are available upon request.

[18] We use the *lpoly* command in Stata to plot the smoothed values of a Kernel-weighted local polynomial regression of the recombination probability on the length of time since the oldest herb was released.

Next, we formally test for the relationship between recombination and the duration of time since an herb's introduction. In the base case, each observation is an herb-patent pair, for all herbs and for all patents. For example, a patent with three herbs would have three herb-level observations. Table 1.4 presents the results of estimating Equation 3. In all specifications, we include herb and year fixed effects in addition to patent-level controls.[19] Column 1 shows a positive and significant relationship between recombination and time elapsed since an herb was introduced (i.e., the difference between the year of application for the focal patent and the first year an herb was used in any patent), with β = 0.0053 and p<0.001. Column 2 further shows an inverted-U-shaped relationship between recombination and time since an herb was introduced (β_1 = 0.0188, p<0.001, β_2 =-0.0004, p<0.0001). We also report estimates for recombination by non-ethnic inventors (Columns 3-4), mixed teams (5-6), and ethnic inventors (7-8). We observe an inverted-U-shaped relationship for recombination by non-ethnic inventors. For mixed teams, we see instead an increasing rate of recombination over time. We formally test for inverted U-shaped relationships as per Simonsohn (2017) and present results in Figure 1.4 (b-c). Intuitively, this method tests whether one observes both a positive and significant slope to the left of a given cutoff and a negative and significant slope to the right, with the cutoff being set via a Robin-Hood algorithm (which the author shows is robust to errors). For non-ethnic teams, recombination increases in the first 16.72 years (β_{low}=0.027, p<0.001) but decreases afterwards (β_{high}= -0.008, p<0.001). For fully ethnic teams, we observe significantly lower levels of recombination; an increase in the first 20.74 years (β_{low}=0.002, p<0.001) and decreases afterwards (β_{high}=-0.003, p<0.001). Mixed teams increase recombination over time and recombination seems to be disproportionately driven by teams with non-ethnic inventors and mixed teams.

[19] We include citation counts, number of claims, number of inventors, assignee age at filing, and whether the patent has new inventors.

Table 1.4. Recombination probabilities over time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable:	Recombined		Recombined (non-ethnic inventors)		Recombined (mixed team)		Recombined (ethnic inventors)	
Time since herb introduced	0.0053 (0.0012)	0.0188 (0.0031)	0.0033 (0.0011)	0.0186 (0.0029)	0.0029 (0.0007)	-0.0027 (0.0020)	-0.0009 (0.0002)	0.0029 (0.0008)
Time since herb introduced squared		-0.0004 (0.0001)		-0.0005 (0.0001)		0.0002 (0.0001)		-0.0001 (0.0000)
Constant	0.1655 (0.0210)	0.0893 (0.0229)	0.2605 (0.0167)	0.1737 (0.0208)	-0.1348 (0.0125)	-0.1033 (0.0137)	0.0398 (0.0043)	0.0188 (0.0057)
Herb FE	Y	Y	Y	Y	Y	Y	Y	Y
Patent Level Controls	Y	Y	Y	Y	Y	Y	Y	Y
Observations	4849	4849	4849	4849	4849	4849	4849	4849
Adjusted R^2	0.020	0.023	0.083	0.088	0.176	0.178	0.062	0.065

Notes: Standard errors appear in parentheses, clustered at the herb level. Each observation is an herb-patent pair, for all herbs and for all patents. For example, a patent with three herbs would count as three data points. We have 758 patents (note that sample size for Table 1.2 is 758) and end up with 4849 herb-patent pairs. The variable "Time since herb introduced" for an herb-patent pair is calculated by subtracting the patent-application year from the first year we observe the focal herb in any patent, across all patents. Our dependent variable measures, for each herb-patent pair, whether the patent uses synthetic compounds. Because we are tracking, for a given herb, the likelihood of recombination over time, we include herb fixed effects. We include patent-level controls to control for patent-level factors that may confound the relationship. The median herb is 17 years old, suggesting that a median herb is 20.4 percentage points more likely to be recombined than a newly introduced herb (from Column (2), 17*0.0188 - 17^2*0.0004=0.3196-0.1156=0.204).

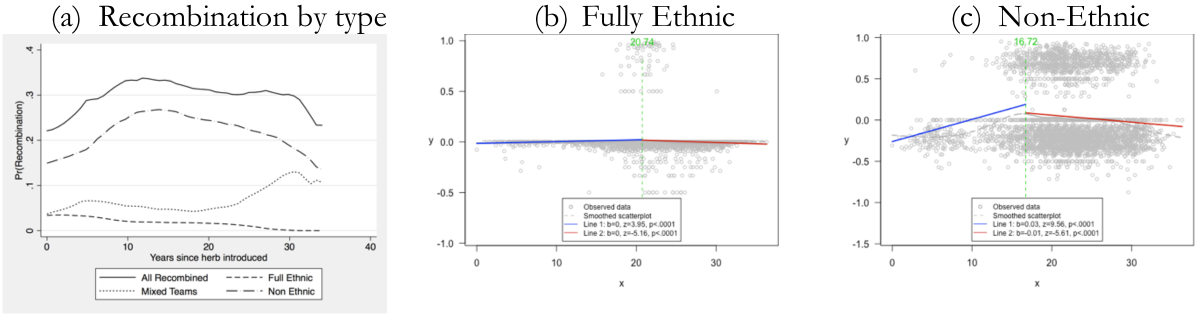

(a) Recombination by type
(b) Fully Ethnic
(c) Non-Ethnic

Figure 1.4. Recombination probabilities for herbs over time since their introduction to the U.S.
Notes: This figure visualizes results reported in Table 1.4 (i.e. recombination probabilities over time). Plots are generated via Kernel-weighted local polynomial regressions using Stata's *lpoly* command. The unit of analysis is at the herb-year level. The x-axis plots the time since an herb was introduced. The dependent variable is whether an herb was used in a patent for recombination at time t. Recombination is measured at the patent level, and can be a recombination using a synthetic compound, along with multiple herbs. Each point on the line thus corresponds to the probability of being recombined (by an ethnic group), for a given herb, t years after its introduction (i.e., $p(t) = E\left[\frac{\gamma_h(t)}{N_h(t)}\right] = \frac{1}{\sum_t N_h(t)} \sum_h \frac{\gamma_h(t)}{N_h(t)}$, where $N_h(t)$ is the number of patents using herb h at time t, and $\gamma_h(t)$ is the number of recombined patents (by an ethnic group) using herb h, at time t. The solid line in panel (a) represents recombination by any ethnic group, and therefore is the sum of all other lines. We see that patents overall and patents by non-Chinese/Indian teams exhibit an inverted-U-shaped pattern with respect to recombination probabilities. In panels (b) and (c), we estimate an interrupted regression with the break point determined by a "Robin Hood" algorithm as in Simonsohn (2017). Panel (b) plots the results for any recombination, and panel (c) plots the results for recombination by fully non-ethnic teams. An inverted U-shape relationship exists if the linear coefficient to the left of the breakpoint is positive and significant, and the linear coefficient to the right of the breakpoint is negative and significant. While not shown here, recombination by mixed teams is gradually increasing over time.

1.6 Robustness checks and secondary analyses

The relevance of the shock to the sample of assignees. An empirical concern is that, if H1B visas were primarily reserved for IT companies, firms that applied for herbal patents might not have been affected by the visa shock. We performed several tests related to determine the relevance of the shock to our sample of firms. Since we cannot observe the number of H1B visa grants at the firm level, we use as a proxy Labor Condition Applications (LCAs, a prerequisite for H-1B visas), which we observe for the universe of firms and thus for each firm in our sample. We match all U.S.-based assignees that were granted herbal patents to company names in LCA filings. Our control group is therefore all non-pharmaceutical companies that filed for LCAs. The control group includes companies such as IBM, Merrill Lynch and Goldman Sachs; herbal patent assignees include firms such as Amgen and Eli Lilly. Figure A2 in the Appendix plots the quantile-quantile plot of total LCAs filed by our assignees and by the control group. The quantile-quantile plot shows that the number of LCAs filed by herbal-patent assignees is left-skewed, suggesting that herbal assignees probably hired more people via H1B visas than did all other firms that filed LCAs. *t*-test results show that herbal assignees filed for 146.7 more LCAs on average (*t*-statistic 4.81) than other firms in the LCA sample, further showing that herbal-patent assignees were indeed a major beneficiary of H1B visas. We also performed back-of-the-envelope calculations of the effect of the visa shock on hiring to show that the H1B shock had a meaningful impact on inventor hiring and on patenting at the treated assignees (cap-subject assignees with herbal patent grants) in our sample. Our calculations suggest that hiring ten additional Chinese/Indian inventors results in 1.605 additional patent grants. This estimate aligns well with, for instance, Amgen in 2015. That year Amgen filed for 420 LCAs, 80 of which led to H1B visas. In 2015, 80% of H1Bs were granted to Chinese/Indians,[20] suggesting that 64 Chinese/Indian nationals working for Amgen received H1B visas. We observe that Amgen filed 71 herbal patents that year, nine of which were granted. This pattern suggests that hiring 10 additional Chinese/Indian inventors results in 1.406 additional herbal patent grants, a number similar to the 1.605 obtained above. Details appear in the Appendix (Section 2).

Placebo test. Serial correlation in the treatment variable across years may bias standard errors in difference-in-differences estimates, causing us to underestimate the standard errors and to over-reject the null hypothesis. We follow the block bootstrapping suggestions in Bertrand, Duflo, and Mullainathan (2004) and Chetty, Looney, and Kroft (2009) to run a nonparametric permutation test to study whether our estimates suffer from such biases. Intuitively, the permutation test calculates the probability that we will see an equally large effect size when the treatment groups and treatment

[20] https://www.recode.net/2017/4/13/15281170/china-india-tech-h1b-visas

periods are randomly selected while preserving the assignee-level correlation structure. Figure A4 in the Appendix plots the cdf of the placebo estimates and a vertical line corresponding to the size of our DD coefficient. As this figure shows, we observe coefficients as large as the one in the fully specified model in Table 1.2 (i.e., 0.044) less than five percent of the time, boosting our confidence in the results.

Inventors' educational backgrounds. An empirical concern about using algorithms that code ethnicities using names is whether the ethnicities of multicultural individuals are identified correctly. Inventors' backgrounds can provide information about whether herbal-patent inventors are more likely to be first-generation migrants. We randomly sampled 552 inventors from the Chinese/Indian population and searched for their educational histories on LinkedIn. To do so, we searched LinkedIn for the inventors' and assignees' names. If we found a profile that (1) matched the inventor name and (2) matched the assignee of interest (3) close in time to the period when the patent application was submitted, we coded a search as successful. We matched 84 profiles (15% of our random sample) but dropped 20 individuals who did not list educational credentials. (See Tables A14–A16 in the Appendix). Approximately one-third of the sample was educated solely in India; a similar fraction was educated solely in the United States. About 20% were educated in China before moving to the United States to study and/or work. The remaining inventors were educated solely in China (9%) or educated in both India and the United States (3%). In summary, a disproportionate fraction of Chinese/Indian inventors who filed herbal patents were educated in China or India, indicating that they are indeed first-generation migrant inventors.

Clustering and alternate specifications. Our analysis pertinent to Hypotheses 1 and 2 is clustered at the assignee level because we believe that error terms for patents with the same assignee will be correlated. For instance, company-specific policies may affect the proportion of ethnic Chinese/Indian inventor names on a given company's herbal patents. An alternative and broader level at which to cluster would be the patents' IPC class, but doing so would reduce the effective number of clusters to <40, which might be too few for unbalanced panels (Cameron and Miller, 2015). Thus, we consider the assignee level appropriate to cluster standard errors. Our results (reported in appendix Table A8) are robust to nonlinear count models (Poisson regression, Poisson QML) and to nonlinear count models that explicitly allow for over-dispersion (negative binomial regressions).

Differences in investments in patent quality. Differences in incentives and resources between capped and exempt assignees raise possible endogeneity concerns. It is possible that capped firms are more likely to benefit from herbal patents, and that they also have greater resources (better knowledge workers and patent lawyers) with which to obtain granted patents. Our use of assignee fixed effects mitigates this concern. In addition, a *t*-test on the different means of grant probabilities between capped and exempt companies returns a difference of 0.039 with a standard deviation of 0.101, suggesting that the grant probabilities of the two groups do not significantly differ.

Value of herbal remedies to the western bio-pharma industry. As for whether herbal remedies are valuable to the western bio-pharma industry, a few stylized facts will shed light. The U.S. herbal-remedies market was worth $5.4 billion in 2016 and is forecasted to reach $6.6 billion by 2021 (Mintel, 2016). Well-known products in this segment include Metamucil (Procter & Gamble), Benefiber (GlaxoSmithKline), and Fibercon (Pfizer) (Euromonitor, 2016). Within western scientific research more generally, herbal and natural ingredients are cited as key sources for drug discovery (Doak *et al.*, 2014). Prior literature documents that, between 1981 and 2014, at least 33 percent of all newly introduced chemical entities (NCEs) were natural-product-derived (Newman and Cragg, 2007). Figure A1 in the Appendix plots counts of articles on herbal remedies in all journals in PubMed and in selective journals like *Science*, *Nature,* and the *New England Journal of Medicine*.

Value created for firms. We next provide evidence that herbal patents filed by first-generation ethnic migrant inventors create value for firms (measured using patent citations). Our secondary analysis (reported in Table A6 in the Appendix) shows that citations of new-herb patents filed by capped firms increased by 91 percent during the visa shock. Capped firms experienced a supply shock of first-generation ethnic migrants during the visa shock. Given our prior finding that patents with new herbs are more likely to be filed by ethnic migrants, this pattern suggests that herbal patents filed by first-generation ethnic migrant inventors create value for firms.

1.7 Discussion and Conclusion

We studied the role of first-generation ethnic migrant inventors in cross-border transfer of knowledge previously locked within the cultural contexts of their home regions. We exploit an exogenous supply shock to U.S. immigration and a list of patenting entities excluded from the shock to present robust econometric results. We also find that ethnic migrant inventors are more likely to engage in reuse of their prior knowledge, whereas knowledge recombination is more likely to be pursued by teams comprising inventors from other ethnic backgrounds.

Contributions of our study

Our results contribute to several literatures, including those on skilled migration, ethnic migration, inventor mobility and knowledge flows, and the microfoundations of knowledge recombination. Like Jensen and Szulanski (2004), who argued that institutional distance increases the stickiness of knowledge and impedes its transfer, we argue that knowledge can be locked within the cultural and linguistic context where it was originally produced. We add to this literature by showing that hiring skilled ethnic migrants can help firms appropriate, and subsequently recombine, knowledge previously locked in the home regions of ethnic migrants.

Our results contribute to the literature on skilled migration. Recent research and policy debate (Kerr and Lincoln, 2010; Kerr *et al.*, 2012; Doran *et al.*, 2016) have focused on the job-creation effects of the H1B program.[21] We sidestep that debate to highlight the role of migrant inventors in transferring across borders knowledge previously locked within their home regions. Contrary to the prevailing assumption that skilled migrants resemble local knowledge workers (and thus might displace them), our paper implies that skilled ethnic migrants can *differ* from locals with regard to the knowledge they bring to a host firm. This finding suggests that debate on skilled immigration should consider knowledge-transfer and knowledge-recombination effects with and without skilled migration. We also contribute to the skilled-migration literature by highlighting the role of first-generation ("new") migrants in knowledge transfer across borders.

There is also an emerging literature on the role of ethnic inventors and Diaspora in facilitating knowledge transfer (Saxenian, 2000; Kerr, 2008; Nanda and Khanna, 2010; Agrawal *et al.*, 2011; Foley and Kerr, 2013; Almeida, Phene, and Li, 2014; Choudhury 2015).[22] We contribute to it by studying the roles of ethnic and non-ethnic inventors in knowledge recombination, a topic that has not been fully explored in the previous literature. Like Freeman and Huang (2014), who find that diversity in author ethnicity is related to more citations and a higher impact factor, our results suggest that knowledge recombination is partly driven by mixed teams.

Our results contribute to the strategy literature on the micro-foundations of knowledge recombination (Allen, 1977; Fleming 2001) by heeding calls to elucidate the micro-foundations of

[21] Kerr and Lincoln (2010) find that changes in H1B admission levels influence the rate of Indian and Chinese patenting in cities and at firms dependent on the program. Kerr *et al.* (2012) find increases in firms' employment of skilled immigrants to be related to overall employment of skilled workers. But Doran *et al.* (2016) find that H1B visas crowd out employment of other workers.

[22] Kerr (2008) notes that ethnic scientific networks are central to short-term technology transfer from the United States. Agrawal *et al.* (2011) find that inventors who work for multinational firms in India cite the Indian Diaspora more frequently than do counterparts employed by the same firms in other countries. Almeida *et al.* (2014) find evidence of intra-ethnic citations in the U.S. semiconductor industry.

innovation within firms (e.g., Felin and Foss, 2005). The recent literature in this area includes the work of Gruber *et al.* (2013) on how inventors' individual characteristics (e.g., their educational backgrounds and whether they are scientists or engineers) affect the breadth of their technological recombinations. Other work (Fleming *et. al.*, 2007; Paruchuri and Awate, 2017) studies the effect of individual inventors' network positions on their ability to engage in recombination.[23] Our findings contribute to this literature by suggesting that, though ethnic migrant inventors may transfer knowledge from their home regions to western firms and reuse it, recombination is apt to be performed by non-ethnic inventors. This scenario indicates a complementary relationship between the ethnic migrant inventor and the non-ethnic inventor, an insight that recalls the literature on concurrent sourcing of complementary components for knowledge recombination (Parmigiani and Mitchell, 2009; Hess and Rothaermel, 2011). [24] Our findings also speak to a mechanism that Nerkar (2003) calls "temporal exploration" whereby firms create value by combining new knowledge with time-honored knowledge.

Generalizability of our results and boundary conditions

External validity is among our study's limitations. To explore the generalizability of our results, we did a comprehensive search of the literature in economic history and migration and profiled nine qualitative examples pertinent to the phenomenon of interest (Table A1 in the Appendix). They include the example of Italian migrants' transfer of knowledge specific to the food and fashion industries to the United States and Australia. As we document, Italians resisted assimilation in general, and specifically resisted "Americanizing" their cooking habits. Italian

[23] Fleming, Mingo, and Chen (2007) study the brokerage positions of individual inventors; Paruchuri and Awate (2017) study the reach of inventors in intra-firm networks and their ability to span structural holes. Other papers in this literature include Nerkar and Paruchuri, 2005; Audia and Goncalo, 2007; and Tzabbar, 2009.

[24] Our results are closely related to those of Hess and Rothaermel (2011), who build on Arora and Gambardella (1990) by arguing that star scientists can link a firm to complementary, non-redundant knowledge at other organizations. Our insights contribute to the broader literature on intra-firm knowledge recombination (Carnabuci and Operti, 2013; Karim and Kaul, 2014).

migrants' transfer of knowledge and ingredients to the United States gave way to subsequent recombination of knowledge: firms such as Campbell's and Heinz marketed shelf-stable versions of Italian cuisine, including such dishes as spaghetti in tomato sauce (Levenstein, 1985). In doing so, the American firms applied their own knowledge of processing and packaging food to classic Italian cooking. We also profile the example of British migrants to Italy who transferred back home their tacit know-how about operating silk machines (Cipolla, 1972). A further example is the transfer of accounting practices by Indian migrants from Gujarat to South Africa.

To reiterate, the key assumption underlying our phenomenon of interest is that, ex-ante, knowledge was initially *locked* in the cultural context of ethnic migrants' home region, due to causal ambiguity, and/or unprovenness of using the knowledge in a different context, and/or concentration in the home region of experts with the requisite tacit know-how. This boundary condition implies that, over time, if both the 'know-why' and the 'know-how' could be codified in the host region, such knowledge could be unlocked and could transfer to a new geography. As an example, Florida and Kenney (1991) study Japanese automotive assembly plants in the United States and conclude that Japanese production practices can be uncoupled from Japanese culture and transferred abroad. Bikard and Marx (2018) have also shown that the same knowledge might be *simultaneously* discovered across two geographical contexts, thus circumventing the need for transfer. Future research could further explore boundary conditions in other settings, such as restaurants (ethnic migrant chefs and sommeliers).

Other limitations and directions for future research

We capture the effects of immigration via the marginal H1B visa candidate, a highly skilled individual. More general increases in immigration may have different impacts on the transfer, codification, and recombination of knowledge previously locked in migrants' home regions.

Beyond the cultural dimension of geographic context, future research can explore how region-specific institutional factors influence why knowledge is locked in a particular geographic context. It can also explore the role in cross-border knowledge transfer of ethnic scientists who are temporary migrants. An example is A.Q. Khan of Pakistan, a temporary migrant scientist who arguably transferred knowledge of centrifuge technology to North Korea.[25] Future work could also study whether ethnic migrants codify such knowledge in forms other than patenting, i.e., academic publications and business practices, and how the value of recombining knowledge from non-western settings with existing western knowledge should be shared between ethnic groups. An example of such a project is the Amazon Third Way initiative, an effort by the World Economic Forum to design and deploy the Amazonian Bank of Codes, an open digital platform that will map the biological assets of the Amazon and provide a global marketplace for such knowledge. In the broader strategy literature, scholars could study the role of skilled ethnic migrants in cross-border transfer of knowledge underlying cultural goods and services.[26]

We began this study by asking whether ethnic migrant inventors differ from locals in how they contribute to innovation at their host firms. Our research outlines at least one dimension along which ethnic migrants differ from non-ethnic knowledge workers, i.e., the knowledge they bring to the firm. Our research suggests that ethnic migrant inventors and non-ethnic inventors play different roles vis-à-vis knowledge reuse and knowledge recombination. Our results have managerial implications for firms engaged in R&D and cross-border sourcing of ideas. They also have implications for policy pertaining to high-skilled immigration and the effectiveness of temporary

[25] Source: https://qz.com/1080927/did-pakistan-help-north-korea-develop-nuclear-weapons-india-us-japan-want-to-know/

[26] There is a rich strategy literature on cultural goods (e.g., Lampel, Lant, and Shamsie, 2000) but little empirical work linking migration of ethnic knowledge workers to the spread of cultural goods across borders.

worker programs like the H1B.[27] In conclusion, our study suggests that a focus on whether migrants create or displace local jobs is too narrowly focused; social planners should consider the loss to cumulative knowledge production if western countries restrict their intake of skilled ethnic migrants.

[27] Source: https://www.wsj.com/articles/indian-workers-in-u-s-fear-trump-h-1b-visa-crackdown-1488191404

References

Agrawal A, Kapur D, McHale J, Oettl A. 2011. Brain drain or brain bank? The impact of skilled emigration on poor-country innovation. *Journal of Urban Economics* **69**(1): 43–55.

Ahuja G, Morris Lampert C. 2001. Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic management journal* **22**(6–7): 521–543.

Allen TJ. 1977. Managing the flow of technology: technology transfer and the dissemination of technological information within the R&D organization, 1st pbk. print. MIT Press: Cambridge, Mass.

Almeida P, Kogut B. 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management science* **45**(7): 905–917.

Almeida P, Phene A, Li S. 2014. The influence of ethnic community knowledge on Indian inventor innovativeness. *Organization Science* **26**(1): 198–217.

Amabile T, Kramer S. 2011. The progress principle: Using small wins to ignite joy, engagement, and creativity at work. Harvard Business Press.

Ambekar A, Ward C, Mohammed J, Male S, Skiena S. 2009. Name-ethnicity Classification from Open Sources. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09. ACM: New York, NY, USA: 49–58. Available at: http://doi.acm.org/10.1145/1557019.1557032.

Arora A, Gambardella A. 1990. Complementarity and external linkages: the strategies of the large firms in biotechnology. *The journal of industrial economics* : 361–379.

Audia PG, Goncalo JA. 2007. Past success and creativity over time: A study of inventors in the hard disk drive industry. *Management Science* **53**(1): 1–15.

Autor DH. 2003. Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of labor economics* **21**(1): 1–42.

Belfanti C. 2004. Guilds, patents, and the circulation of technical knowledge: Northern Italy during the early modern age. *Technology and culture* **45**(3): 569–589.

Berry H, Guillén MF, Zhou N. 2010. An institutional approach to cross-national distance. *Journal of International Business Studies* **41**(9): 1460–1480.

Bertrand, M., Duflo, E., Mullainathan, S. 2004. How much should we trust differences-in-differences estimates?. *The Quarterly journal of economics*, **119**(1): 249-275.

Bikard, M. and Marx, M., 2018. Hubs As Lampposts: Academic Location and Firms' Attention to Science.

Borjas GJ, Doran KB. 2012. The collapse of the Soviet Union and the productivity of American mathematicians. *The Quarterly Journal of Economics* **127**(3): 1143–1203.

Breschi S, Lissoni F. 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of economic geography* **9**(4): 439–468.

Burt RS. 1992. *Structural holes: the social structure of competition*. Harvard University Press: Cambridge, Mass.

Cameron AC, Miller DL. 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* **50**(2): 317–372.

Cameron AC, Trivedi PK. 2010. *Microeconometrics using stata*. Stata press College Station, TX, 2.

Carnabuci G, Operti E. 2013. Where do firms' recombinant capabilities come from? Intraorganizational networks, knowledge, and firms' ability to innovate through technological recombination. *Strategic Management Journal* **34**(13): 1591–1613.

Chetty R, Looney A, Kroft K. 2009. Salience and taxation: Theory and evidence. *American economic review* **99**(4): 1145–77.

Choudhury P. 2015. Return migration and geography of innovation in MNEs: a natural experiment of knowledge production by local workers reporting to return migrants. *Journal of Economic Geography* **16**(3): 585–610.

Cipolla CM. 1972. The diffusion of innovations in early modern Europe. *Comparative Studies in Society and History* **14**(1): 46–52.

Cohen WM, Levinthal DA. 1990. Absorptive capacity: a new perspective on learning and innovation. *Administrative science quarterly* : 128–152.

Connelly BL, Certo ST, Ireland RD, Reutzel CR. 2011. Signaling theory: A review and assessment. *Journal of management* **37**(1): 39–67.

Cowan R, Foray D. 1997. The economics of codification and the diffusion of knowledge. *Industrial and corporate change* **6**(3): 595–622.

Dasgupta P, David PA. 1994. Toward a new economics of science. *Research policy* **23**(5): 487–521.

Doak BC, Over B, Giordanetto F, Kihlberg J. 2014. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chemistry & biology* **21**(9): 1115–1142.

Doran K, Gelber A, Isen A. 2016. The effects of high-skilled immigration policy on firms: Evidence from H-1B visa lotteries. National Bureau of Economic Research.

Euromonitor. 2016. Herbal/traditional products in the US. Available at: http://www.portal.euromonitor.com.ezp-prod1.hul.harvard.edu/portal/analysis/tab [15 March 2017].

Felin T, Foss NJ. 2005. *Strategic organization: A field in search of micro-foundations*. Sage Publications London, Thousand Oaks, CA and New Delhi.

Fleming L. 2001. Recombinant uncertainty in technological search. *Management science* **47**(1): 117–132.

Fleming L, Mingo S, Chen D. 2007. Collaborative brokerage, generative creativity, and creative success. *Administrative science quarterly* **52**(3): 443–475.

Florida R, Kenney M. 1991. Organisation vs. culture: Japanese automotive transplants in the US. *Industrial Relations Journal* **22**(3): 181–196.

Foley CF, Kerr WR. 2013. Ethnic innovation and US multinational firm activity. *Management Science* **59**(7): 1529–1544.

Franzoni C, Scellato G, Stephan P. 2014. The mover's advantage: The superior performance of migrant scientists. *Economics Letters* **122**(1): 89–93.

Freeman RB, Huang W. 2014. Collaboration: Strength in diversity. *Nature News* **513**(7518): 305.

Ganguli I. 2015. Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *Journal of Labor Economics* **33**(S1): S257–S288.

Ghemawat P. 2001. Distance still matters. *Harvard business review* **79**(8): 137–147.

Gruber M, Harhoff D, Hoisl K. 2013. Knowledge recombination across technological boundaries: Scientists vs. engineers. *Management Science* **59**(4): 837–851.

Hambrick DC, Macmillan IC. 1985. Efficiency of product R&D in business units: The role of strategic context. *Academy of Management Journal* **28**(3): 527–547.

Henderson RM, Clark KB. 1990. Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly* : 9–30.

Hess AM, Rothaermel FT. 2011. When are assets complementary? Star scientists, strategic alliances, and innovation in the pharmaceutical industry. *Strategic Management Journal* **32**(8): 895–909.

Holmström B. 1999. Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies* **66**(1): 169–182.

Hornung E. 2014. Immigration and the diffusion of technology: The Huguenot diaspora in Prussia. *American Economic Review* **104**(1): 84–122.

Jensen R, Szulanski G. 2004. Stickiness and the adaptation of organizational practices in cross-border knowledge transfers. *Journal of international business studies* **35**(6): 508–523.

Johnson CY. 2017, February 1. Big pharma depends on immigrants. It kept quiet about Trump's travel ban. *Washington Post*. Available at: https://www.washingtonpost.com/news/wonk/wp/2017/02/01/big-pharma-depends-on-immigrants-it-kept-quiet-about-the-travel-ban/.

Karim S, Kaul A. 2014. Structural recombination and innovation: Unlocking intraorganizational knowledge synergy through structural change. *Organization Science* **26**(2): 439–455.

Kerr SP, Kerr W, Özden Ç, Parsons C. 2016. Global talent flows. *Journal of Economic Perspectives* **30**(4): 83–106.

Kerr SP, Kerr WR, Lincoln WF. 2015. Skilled immigration and the employment structures of US firms. *Journal of Labor Economics* **33**(S1): S147–S186.

Kerr WR. 2008. Ethnic scientific communities and international technology diffusion. *The Review of Economics and Statistics* **90**(3): 518–537.

Kerr WR, Lincoln WF. 2010. *The supply side of innovation: H-1B visa reforms and US ethnic invention.* National Bureau of Economic Research. Available at: http://www.nber.org.ezp-prod1.hul.harvard.edu/papers/w15768.

Lampel J, Lant T, Shamsie J. 2000. Balancing act: Learning from organizing practices in cultural industries. *Organization science* **11**(3): 263–269.

Levenstein H. 1985. The American response to Italian food, 1880–1930. *Food and Foodways* **1**(1–2): 1–23.

Lewin AY, Massini S, Peeters C. 2009. Why are companies offshoring innovation? The emerging global race for talent. *Journal of International Business Studies* **40**(6): 901–925.

Michel S, Witte J. 2014. *Immigrants Working for U.S. Pharmaceuticals.* George Mason University, Fairfax, VA. Available at: http://s3.amazonaws.com/chssweb/documents/16298/original/Immigrants_in_the_Pharma_Industry_Institute_for_Immigration_Research_GMU.pdf?1407181243.

Mintel. 2016. Homeopathic and Herbal Remedies - US - November 2016 - Market Research Report. Available at: http://academic.mintel.com.ezp-prod1.hul.harvard.edu/display/765799/ [15 March 2017].

Nanda R, Khanna T. 2010. Diasporas and domestic entrepreneurs: Evidence from the Indian software industry. *Journal of Economics & Management Strategy* **19**(4): 991–1012.

Nelson RR, Winter SG. 1982. *An Evolutionary Theory of Economic Change.* SSRN Scholarly Paper, Social Science Research Network, Rochester, NY. Available at: https://papers.ssrn.com/abstract=1496211.

Nerkar A. 2003. Old is gold? The value of temporal exploration in the creation of new knowledge. *Management Science* **49**(2): 211–229.

Nerkar A, Paruchuri S. 2005. Evolution of R&D capabilities: The role of knowledge networks within a firm. *Management science* **51**(5): 771–785.

Newman DJ, Cragg GM. 2007. Natural products as sources of new drugs over the last 25 years. *Journal of natural products* **70**(3): 461–477.

Oettl A, Agrawal A. 2008. International labor mobility and knowledge flow externalities. *Journal of international business studies* **39**(8): 1242–1260.

Parmigiani A, Mitchell W. 2009. Complementarity, capabilities, and the boundaries of the firm: the impact of within-firm and interfirm expertise on concurrent sourcing of complementary components. *Strategic Management Journal* **30**(10): 1065–1091.

Paruchuri S, Awate S. 2017. Organizational knowledge networks and local search: The role of intra-organizational inventor networks. *Strategic Management Journal* **38**(3): 657–675.

Polanyi M. 1966. *The tacit dimension*, 1st ed. Terry lectures. Doubleday: Garden City, N.Y.

Rosenkopf L, Almeida P. 2003. Overcoming local search through alliances and mobility. *Management science* **49**(6): 751–766.

Saxenian A. 2000. Silicon Valley's New Immigrant Entrepreneurs. Available at: https://escholarship.org/uc/item/88x6505q.

Schumpeter JA. 1939. *Business cycles*. McGraw-Hill New York, 1.

Simonsohn U. 2017. Two-lines: The first valid test of U-shaped relationships.

Singh J. 2005. Collaborative networks as determinants of knowledge diffusion patterns. *Management science* **51**(5): 756–770.

Song J, Almeida P, Wu G. 2003. Learning–by–Hiring: When is mobility more likely to facilitate interfirm knowledge transfer? *Management Science* **49**(4): 351–365.

Szulanski G. 1996. Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic management journal* **17**(S2): 27–43.

Szulanski G. 2002. Sticky knowledge: Barriers to knowing in the firm. Sage.

Tzabbar D. 2009. When does scientist recruitment affect technological repositioning? *Academy of Management Journal* **52**(5): 873–896.

Von Hippel E. 1994. "Sticky information" and the locus of problem solving: implications for innovation. *Management science* **40**(4): 429–439.

Vroom VH. 1964. *Work and motivation.* Wiley: Oxford, England.

2. Product Market Performance and Openness: the Moderating Role of Customer Heterogeneity

Do Yoon Kim

Abstract

Increasingly, a firm's ability to create value depends on its ability to achieve alignment with its ecosystem of suppliers and complementors. In such contexts, a firm's strategy may encompass "opening" some parts of their intellectual property (IP) to facilitate value creation, but potentially at the cost of decreased value capture. In this paper, I study the role of customer heterogeneity as a determinant of firms' value capture. I hypothesize that openness creates complement innovations which creates value for customers. These complements are valued by a subset of customers who importantly can mitigate information imperfections in the product market. I further hypothesize such strategies will be more effective for those products that face greater information imperfections. I test these hypotheses on a unique dataset of wireless routers characteristics, complement innovation (custom firmware for wireless routers) compatibility, and product reviews. I utilize an exogenous shock to complement innovation compatibility from several exogenous "reverse engineering" events. I find that the availability of complement innovations increases review ratings by 0.67 stars. I find a strong sorting effect: customers who are more likely to use the custom firmware leave more positive reviews. Importantly, there is a strong information effect as well: these users provide more helpful reviews. Consistent with the information imperfections framework, I find that such effects are stronger for enterprise products which face greater uncertainty.

2.1 Introduction

Increasingly, a firm's ability to create value depends on its ability to achieve alignment with its ecosystem of suppliers and complementors. Across various high technology industry settings, the value delivered to the customer is determined not by the singular effort of a focal firm, but by the coordinated efforts of multiple firms (Bresnahan and Greenstein 1999, Kapoor and Furr 2015, Macher and Mowery 2004). In such interrelated settings, firms must consider not only their strategic positioning, but comprehensive strategies for value creation, capture, and competition in these industries (Brandenburger and Nalebuff 2011). Central to this literature are the distinct organizational, technological, and competitive challenges faced by companies situated in ecosystem contexts.

In such contexts, a firm's strategy may encompass "opening" some parts of their intellectual property (IP). Across multiple industries, media reports firms embracing open IP, from the computer industry (Chan 2018, Metz 2015a, b), to the automotive industry (Musk 2014), and the pharmaceutical industry (Butler 2010). In many cases, articles highlight the importance of this in furthering ecosystem innovativeness. Such notions are echoed in a growing literature that documents how varying degrees of openness can attract complementors and their innovations (Boudreau 2010, Gawer and Henderson 2007, Parker and Van Alstyne 2017, Wen et al. 2015).

This paper studies how opening IP affects value capture and value creation in ecosystem contexts. In particular, I show opening IP can create complementary goods that customers can use in conjunction with the focal product, benefiting them. Furthermore, complement goods can serve to reduce information imperfections, and thus differentially affecting firms' competitive advantage (Oberholzer-Gee and Yao 2018). One way in which complement innovations reduce information imperfections is through attracting technologically "savvy" customers, and through information spillovers to less-savvy customers. I hypothesize that customers vary in their absorptive capacity, driving savvy customers to adopt the complement innovation. This, in turn, informs less-savvy customers about the intangible qualities of the good. I also hypothesize that the reduction of information imperfection should be greater for more complex products for which existing imperfections were significant.

The empirical context in this paper is the wireless router industry, with "custom firmware" as the focal downstream complementary good. In 2003, Cisco was found to be in violation of the GNU General Public License (GPL)[28], and was forced to release the source code for its WRT54G, a popular wireless router product (Meeker 2005). A community of hackers formed around the source code with the stated goal of creating an extendible operating system complete with packages[29]. These software projects, called "custom firmware" projects, allowed customers to enhance their routers' capabilities to better suit their needs (Pash 2006). One of the projects, OpenWRT, has grown significantly, with over 500 individual contributors and 40,000 commits.

A key identification challenge in estimating the impact of downstream complements on performance is the endogeneity surrounding the anticipation and subsequent adoption of the technologies by focal firms. For example, once user communities around downstream complementary goods have formed, focal firms may find it optimal to adopt the complementary product's features into their own. This would allow focal firms to outsource product development to communities. Such incentives would bias any estimates of the value of downstream complements. To address this issue, I exploit a series of events which allowed some products to benefit from custom firmware, independently of manufacturer intent. I document four events in which the "device driver", a crucial software component that allows users to run the complementary software,

[28] The GPL is the prime example of a "reciprocal" license in which distributors of source code are legally obligated to make the source code available to the public.

[29] https://openwrt.org/

was reverse engineered. Without reverse engineering, device drivers were kept proprietary for legal ambiguities and strategic reasons (Henkel 2006, Henkel et al. 2014).

To test my hypotheses, I collect a unique dataset of wireless routers. I combine data from three sources: an online database of wireless router releases and hardware characteristics, open source software repositories, and customer review data (McAuley et al. 2015). The unique dataset allows me to distinguish the hardware and technology suppliers (e.g., Broadcom, Intel, Qualcomm), the focal firms, or original equipment manufacturers (e.g., ASUS, Linksys, Belkin), and product performance (e.g., ratings, review counts). Furthermore, for each product, I am able to pinpoint the time at which it became compatible with the downstream complementary software, allowing me to compare product performance before and after the downstream complement emerges.

In the first set of results, I document how the emergence of OpenWRT and custom firmware, the complementary good of interest, created value for customers on average. I utilize a differences-in-differences approach to estimate the impact of open source on ratings and sales. Comparing the average review ratings for products in the "treatment group" to products that are never compatible with open source, I find open source software increases product ratings by 0.63 stars, an 18 percent increase compared to the pre-period. The effects are robust to controlling for reviewer characteristics, as well as review text.

Second, I examine the role of customer heterogeneity in driving these positive effects. Specifically, I test whether savvy customers have an information effect. I show that "savvy" customers' reviews are more helpful to other users; that "savvy" users are attracted by complement innovations; and also savvy users value complement innovations. For customers that have below median costs (above median savviness), custom firmware increases review ratings by 0.6 stars. For customers with above median costs (below median savviness) I do not find increased review ratings. Furthermore, I find evidence that the presence of complementary goods attracts customers with low adoption costs. Reviews for products compatible with OpenWRT are 13.4 percent more likely to attract a customer with below median costs (above median savviness).

Finally, I document how the complexity of the product may moderate these effects. Within routers, "enterprise" class wireless routers provide better hardware capabilities, thus enhancing the benefits to custom firmware. I categorize routers as enterprise routers if the description contains keywords such as "enterprise" or "business", and all other routers as consumer routers. Using the number of reviews as a proxy for product market performance, I find that custom firmware increases product market performance for enterprise routers.

This contributes to the larger literature on ecosystems and platform organizations (Adner and Kapoor 2010, Evans et al. 2008, Jacobides et al. 2018, Parker and Van Alstyne 2017). In particular, this paper highlights the role of customers in determining the benefits from complementary goods. In doing so, it reinforces the importance of "alignment" between a firm, its customers, and its suppliers as an important determinant of value creation. This paper thus complements the innovation ecosystem literature and provides support for the larger body of theoretical work on ecosystems (Adner 2017, Jacobides et al. 2018). The results should be useful to other researchers in the area in understanding the factors that impact alignment within the ecosystem.

I also contribute to the open source literature by distinguishing between corporate use of open source, and end users' use of open source. This complements literature on firms' strategic use of open source software (Henkel et al. 2014, Nagle 2015, 2018, Wen et al. 2015). In my setting, firms can choose to develop products using open source. Post-production, users who buy the product can replace the existing software with custom versions of open source software. For instance, this allows users to customize and extract greater utility out of their routers (Pash 2006). Firms' use of open source, on the other hand, allows for code reuse from project to project, and can serve to reduce development costs.

In what follows, I present theory and hypotheses in Section 2, followed by Section 3 where I discuss my empirical setting and identification. Section 4 describes data collection and variable definitions, Section 5 outlines the empirical specifications, and Section 6 presents the results. I conclude in Section 7.

2.2 Theory and Hypotheses

Generally, openness involves a trade-off between increased value creation at the cost of decreased value capture and risk of imitation (Alexy et al, Boudreau 2010, West 2003). In this section, I first summarize the literature on openness and lay out this tradeoff in more detail. I then introduce the market imperfections framework (Oberholzer-Gee and Yao 2018) and hypothesize how opening IP reduces market imperfections by mobilizing a subset of users to reduce information asymmetries. I define opening IP as waiving the right to exclude third parties from using an upstream component. Opening IP involves both devolving control and granting access (Boudreau 2010), but in this setting is closer to devolving control.

2.2.1 Opening IP and complementary product innovations

Anecdotal evidence of firms opening up IP spans multiple industries. While most prevalent in the software industry (Chan 2018, Mannes 2017, Metz 2015a, b), it also exists in the automotive industry (Golson 2014, Musk 2014), as well as the pharmaceutical industry (Boseley 2010, Butler 2010). In many cases, firms state “innovation” as the main driver for opening. Microsoft’s decision to open patents was “to encourage innovation with open source software” (Andersen 2018).

Similarly, Tesla and Toyota's decisions to open their patents intended to spur innovation in the electric vehicle industry (Hu et al. 2017, Musk 2014, Toyota 2015, Undercoffler 2015).

Prior work documents how opening IP increases complement innovation. I follow Adner and Kapoor (2010) in defining complements as "goods combined with the focal product downstream by users." For example, opening IP led to innovations in complementary hardware for handheld devices (Boudreau 2010), encouraged open source entry by allowing entrants to evade litigation threats (Wen et al. 2015), and general innovation in the semiconductor industry (Gawer and Henderson 2007). Researchers also find that the effects on complement innovation are moderated by the degree of openness (Laursen and Salter 2006; Boudreau 2010), as well as the duration of protection (Parker & Van Alstyne 2017).

Complement innovations, in turn, should positively affect value created for customers through option value and complementarities. First, note that complement innovations cannot negatively impact the value to the customer. Because the focal product functions independently, the complement innovation is non-unique complements (Jacobides, Cennamo and Gawer), and users can always choose to not adopt the complement. This endows the complement with optionality (Baldwin and Clark). For example, in the wireless router market, customers can choose to use standalone routers, or to enhance the software capabilities by installing third-party firmware. Second, the extent to which the value is positively affected will depend on the complementarities between the focal product and the complement innovation. Jacobides, Cennamo, and Gawer (2016) suggest three types of complementarities. Of these, I am concerned with complement innovations that are non-unique and specialized.

Opening IP, however, increases imitation and thus competition, limiting how much value a single firm can capture (Alexy et al 2018; West 2003). The literature has been largely theoretical, exploring how firms are able to navigate the ease of imitability through mixing proprietary and open components, as well as through job market signaling (Casadesus-Masanell and Llanes 2011, 2015, Kumar et al. 2011). In contrast to the "horizontal" strategies listed before, recent work has shown there exist "vertical" strategies in which firms may choose to open technologies to replace their inputs (Gambardella and Von Hippel 2019). In these studies, the benefits to openness depend on the relative quality of the modules, on the ability of engineers to learn, and the relative efficiency of the downstream firms to upstream suppliers. However, empirical studies of the phenomenon are limited.

2.2.2 Complement innovations, information imperfections, and customer value

An alternative framework to analyze the impact of openness on firm performance is through analyzing market imperfections (Oberholzer-Gee and Yao 2018; Yao 1988). The market imperfections framework argues that competitive advantage arises from frictions that prohibit perfect competition. Three forms of market imperfections are highlighted: production economies, information imperfections, and transaction costs. Each market imperfection acts to soften competition and provide sustainable competitive advantage.

Significant amounts of market imperfections exist in the wireless router industry. Of particular interest are information imperfections. Information imperfections are defined as "Buyer inability to assess quality, or buyer or seller inability to observe relevant actions of other party." In the case of wireless routers, because software is an experience good, sellers ex-ante may have more information about product quality than buyers, despite detailed hardware specifications. Furthermore, sellers may ex-post issue software updates that harm the customer, for instance by collecting user information and selling it to potential buyers[30].

[30] https://www.wsj.com/articles/hundreds-of-thousands-of-routers-are-being-primed-for-a-cyberattack-1527110611 https://www.wired.com/2017/02/smart-tv-spying-vizio-settlement/

Opening IP reduces information imperfections in two ways. First, opening IP reduces ex-ante information asymmetry between buyers and sellers. Under open source software licenses, users are free to study the source code and understand which components go into building the software. This information can reduce information asymmetries in the pre-purchasing stage, which is important for software which is generally an experience good. Furthermore, many open source licenses are copyleft licenses under which distribution of a product mandates the joint distribution of source code. Copyleft licenses may act to deter sellers' malicious intent through increased monitoring, reducing ex-post information asymmetries (Baker, Gibbons, Murphy). For instance, open source is often cited as a pre-requisite to deter data harvesting and other privacy issues.

Taken in conjunction with increased complement innovations, opening IP leads to an increase in customer valuation of the product. The increase will depend on the complementarities between the focal product and the complement, but is non-negative because users can always choose not to adopt the complement. Opening IP further decreases information asymmetries, decreasing search costs for customers and allowing buyers to monitor the sellers ex post. This brings me to my first hypothesis:

H1: Opening IP leads to complement innovations, increasing perceived value

2.2.3 Role of customer heterogeneity in reducing information imperfections

In addition to the direct effect of openness on market imperfections, I argue customer level heterogeneity is a driver of this mechanism. I first introduce the context of marketplaces for reviews, showing they reduce information asymmetries. Next, I build on prior literature to hypothesize the existence of a "savvy" subset of users for whom the buyer-seller information asymmetries are less pronounced. These savvy users are attracted by complement innovations, and aid other customers' purchase decisions.

Reductions in information imperfections can be further facilitated by marketplaces for reviews. Many markets now provide customer reviews, providing potential purchasers with additional information about the product (Mayzlin et al 2014, Faulhaber and Yao 1989; Lee and Hosanagar 2016). Such reviews are generally construed to reduce information asymmetries in the market (Chen and Xie; Faulhaber and Yao 1989).

I define a user (buyer) as "savvy" if he or she possesses information about a product that other buyers will find useful. The literature is abound with examples of such customers. For instance, lead users generally possess needs that precede the needs of other buyers in the market, suggesting they are more attuned to qualities of the product that matter (Urban and Von Hippel 1988; von Hippel 1988; Franke et al 2006; Schreier et al 2007). In particular, these lead users perceive new technologies as "less complex" and lead opinions (Schreier et al 2007). In the case of open source software, companies are able to learn from the crowd of users to capture value, further providing evidence that savvy users exist (Nagle 2018).

Savvy users may have privileged information due to their absorptive capacity. Savvy users, on the other hand, possess more absorptive capacity than their peers (Cohen and Levinthal 1990). In contrast to the general conception of absorptive capacity as a firm level construct, here I focus on the individual level of absorptive capacity. Users with more absorptive capacity allows them to absorb more "sticky" knowledge, that are closer to the locus of innovation. Thus, they are able to extract more value from complement innovations, should they arise, and can adopt these innovations more freely. Note that savvy users in my context are distinct from lead users in that they need not anticipate demand for products.

Savvy users can play an important role in solving information imperfections by providing helpful information to others. A significant literature documents how online marketplaces can reduce information asymmetries (Faulhaber and Yao, 1989; Bakos 1997; Lizzeri 1999). Research on review markets (in particular, online review systems) highlights the link between aggregate sales and reviews, and also how reviews influence individual purchasing decisions[31]. For instance, online reviews reduce search costs and reduce product uncertainty (Anderson, 2008; Bergmann and Ozmen, 2006). Reviews may increase sales (Chen et al 2008; Chevalier and Mayzlin 2006), but the helpfulness of reviews matters (Chen et al 2008). Not all reviews are created equal, as reviews that are longer and more in-depth are also more helpful (Mudambi and Schuff, 2010).

Within such review platforms, users' backgrounds are an important factor in determining the informativeness of reviews. Prior work has shown that disclosing the identity of the reviewer leads to positive ratings of reviews and sales (Forman, Ghose, and Wiesenfield, 2008). Furthermore, lead users are more likely to lead opinions rather than seek opinions (Schreier et al 2007). Finally, there may be heterogeneous preferences across early adopters and late reviewers (Li and Hitt, 2008). This brings me to my second hypothesis:

H2: Savvy users provide more informative reviews, and value and are attracted to products with complement innovations

2.2.4 Strategic implications

Thus far, I have hypothesized that opening IP leads to complement innovation that will be valued by savvy users and attract them. In this section, I hypothesize how strategies of opening IP

[31] For a comprehensive literature review, please read King, Racherla, and Bush (2014) and Cheung and Thadini (2012).

can have heterogeneous implications across firms. Specifically, I argue that given the role of savvy users in mitigating informational asymmetries, the effect will vary across products' levels of information imperfections. I propose directly testing savvy users' review helpfulness, as well as performance implications.

Information asymmetry between the seller and buyer may vary across different markets and different products. For many products, the buyer is unable to assess the quality of the product until after the sale (Akerlof 1970; Arrow 1962). Empirical results support this distinction and find that in the case of review systems, hedonic goods benefit more from other reviewers reviews, presumably because of incomplete information (Lee and Hosanagar, 2016).

In particular, complex products and infrequent purchases may be more susceptible to information asymmetry. Relevant to my setting, I hypothesize that enterprise routers may suffer from greater levels of information imperfections. While routers are physical products whose attributes are easily discerned, the software component of routers are less easy to discern. Furthermore, for enterprise products, strong existing buyer-seller networks are likely to exist. The products are more expensive, and thus buyers face greater uncertainty. Thus, there will be a different informational effect of savvy customers' sorting and reviewing enterprise and non-enterprise products.

H3: Complement innovations will lead to more helpful reviews by savvy customers and enhance product sales, particularly for enterprise products

2.3 Setting and Identification Strategy

This section provides details regarding the industry setting and the identification strategy. In section 3.1, I describe wireless routers and the industry background. In section 3.2, I detail the events leading up to the emergence of an open source operating system and an exogenous increase in compatibility.

2.3.1 Setting: Wireless router industry

To estimate the impact of openness on product market performance, I study the wireless router industry. Owing to the ubiquity of wireless communication, the industry has grown significantly in recent years. In 2017, sales of wireless local area network (WLAN) equipment reached 5.7 billion dollars[32]. It is a particularly interesting industry to study the impacts of openness on firm performance because of its unique history regarding open IP. Due to events which I describe in detail below, the industry evolved from a closed to a more open equilibrium, with companies forming consortia to promote the use of open source software within the industry[33].

Wireless routers are comprised of interacting hardware and software components. In particular, the software exists as a spectrum, as seen in Figure 2.1. At the leftmost end exists the hardware components (e.g., CPU, antennae, LED lights) and at the rightmost end exist the software components that users directly interact with (e.g., applications, terminal, graphical user interface). Between the hardware and the user-facing software, there are three intermediate levels of software: firmware, device drivers, and operating systems. Firmware is software that usually exists within the hardware itself, extending the hardware capabilities. Device drivers provide a unified software-based interface to the hardware to be used by the operating system. Finally, the operating system manages and allocates resources to the user-facing software.

[32] Gartner 2018

[33] https://prplfoundation.org/

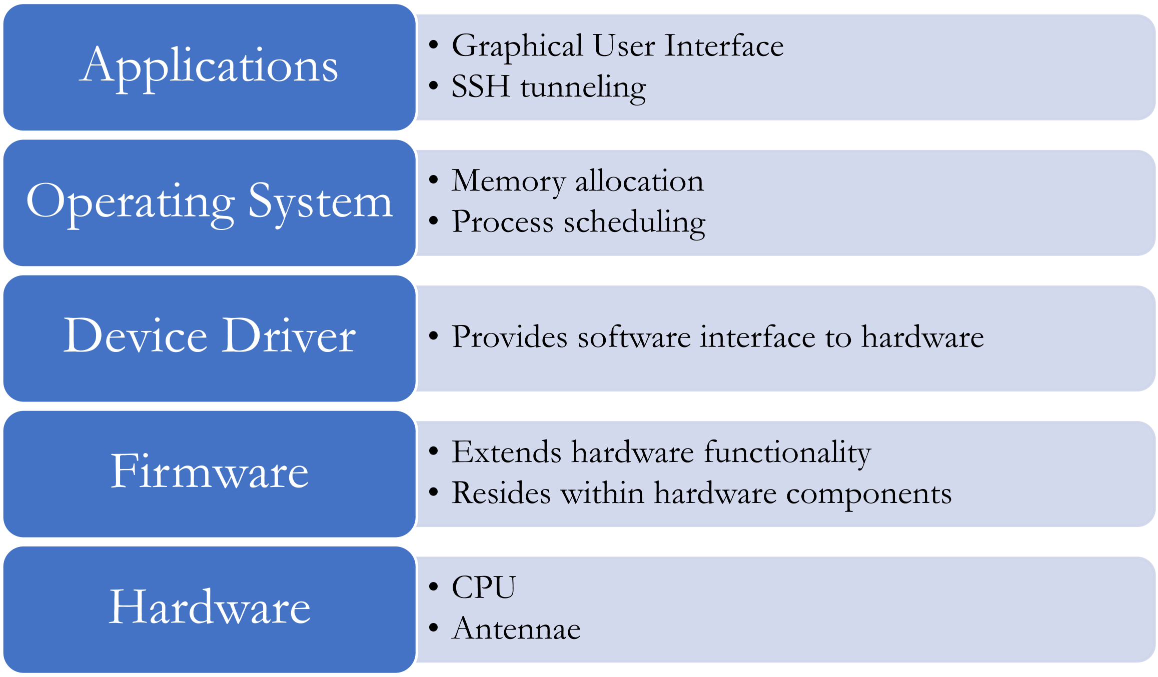

Applications
• Graphical User Interface
• SSH tunneling
Operating System
• Memory allocation
• Process scheduling
Device Driver
• Provides software interface to hardware
Firmware
• Extends hardware functionality
• Resides within hardware components
Hardware
• CPU
• Antennae

Figure 2.1. Router internal components

The "device driver" is a fundamental software component that serves as a bridge connecting the focal product (router) and the complementary good (OpenWRT). The device driver provides a standard interface with which the operating system communicates with the hardware (Corbet et al. 2005). For instance, device drivers translate the operating system's command to turn on wireless connectivity into instructions at the hardware level. They are modular pieces of code, serving as black boxes that allow a particular piece of hardware respond to a well-defined internal programming interface, hiding the details of how the device works.

In the context of wireless routers, the intellectual property that is opened is generally the software components. Under closed software IP, the router manufacturer has the sole authority to create software for its products. Once IP is opened, the router manufacturers can no longer exclude other developers from replacing and extending the software.

Software openness can vary by its location. Device drivers can be hidden because they are in a legal gray area for GPL violations (Henkel 2014). In part, there are legal reasons for not releasing low level code that allows users to tweak the radio frequencies[34], but there also are strategic considerations (Henkel 2006). Thus, when Cisco/Linksys released the source code for the Linksys WRT54G, they released the device driver in binary form: doing so would allow compatibility with the GPL, but disallow users from tweaking the radio frequencies.

2.3.2 Complement innovations in the wireless router industry: Free and open source "custom firmware"

In December of 2002, Linksys (a wireless router OEM) released the WRT54G, a popular wireless router. In 2003, Linksys was in the process of being acquired by Cisco when the Free Software Foundation demanded source code for the WRT54G. Earlier that year, Andrew Miklas posted to the Linux Kernel Mailing List, a popular email list for Linux developers, that Cisco is distributing Linux based products without the corresponding source code. The Free Software Foundation stepped in and acted as a mediator between the community, Cisco/Linksys, and Broadcom. After some negotiation, the source code was released in late 2003[35]. This attracted the attention of many programmers worldwide, and led to one of the most successful free and open source software projects for embedded systems: OpenWRT. As of 2018, the main OpenWRT project had received 42,463 commits from 419 contributers, and 4,525 packages (or apps) from 416 contributors. The success of OpenWRT led to similar projects such as DD-WRT, AsusWRT, and Tomato.

[34] For example, the FCC prohibits users from modifying the radio frequency (RF) outside of its approved range (https://apps.fcc.gov/kdb/GetAttachment.html?id=zXtrctoj6zH7oNEOO6De6g%3D%3D&desc=594280%20D02%20U-NII%20Device%20Security%20v01r03&tracking_number=39498)

[35] Note this event, while related, is distinct from the legal events surrounding Free Software Foundation, Inc. v. Cisco Systems, Inc. (https://en.wikipedia.org/wiki/Free_Software_Foundation,_Inc._v._Cisco_Systems,_Inc.). After the events covered in this paper, Cisco continued to violate the GPL in other products it released. The FSF filed a suit against Cisco (the only lawsuit filed by FSF), but the case was settled out of court.

Users can replace some of the software components with this custom firmware, the focal downstream innovation in this study. Simply by downloading and installing software onto the router, users are able to enhance their routers' capabilities. In addition to users benefiting from tinkering with the software, it sometimes allows users to take advantage of software functions that are only available in higher end routers[36].

2.3.3 Identification Strategy

The ideal experiment to test the broader question of the strategic role of opening IP on firm performance would involve randomizing IP openness across multiple industries and measuring firm profits in the wake of such changes. Such experiments would allow the findings to generalize to other settings and allow researchers to uncover the moderating effects of industry structure, but would be prohibitively difficult to achieve. Alternatively, I document a setting in which opening IP led to the creation of a complementary good. This allows me to document the impact of complementary goods on product market performance, a crucial step towards understanding the strategic role of opening IP.

While the emergence of OpenWRT in the wireless router industry was largely unanticipated, once it was introduced, corporations had the choice to embrace or block such developments. Cisco/Linksys tried to block users from utilizing such tools. For instance, in 2005, Cisco/Linksys released an upgraded version of the WRT54G (Corbet 2005), but with smaller ram and flash so as to prevent users from installing custom firmware (Cassia 2006)[37]. Similarly, TP-Link tried to ban custom firmware (Brodkin 2016). On the other extreme, some firms market their products by

[36] https://lifehacker.com/turn-your-60-router-into-a-600-router-178132

[37] In this case, the community was able to get the WRT54G working with Linux, but in many cases this is impossible as stated on the DD-WRT website (https://wiki.dd-wrt.com/wiki/index.php/Known_incompatible_devices)

explicitly highlighting compatibility with custom firmware. Upstream, many large semiconductor companies provide OpenWRT as the central development platform[38].

As such, simple comparisons of means in product performance for open source and non-open source products will be biased by firms' strategic intent. Even controlling for product characteristics through fixed effects, differences may be biased at the component level (i.e., by router companies' selection of SOCs that have OpenWRT).

As stated above, the absence of source code for device drivers restricted routers' compatibility with open source operating systems. One legal way around binary drivers is to reverse engineer the software code to ensure FCC compliance. This usually involves a "clean room" reverse engineering approach, where one group of programmers write up the documentation for how the device driver works, and another group of programmers code the software.

I collect historical data on 4 reverse engineering events (Broadcom, Atheros, Texas Instruments, and Prism) and a list of 25 SOC models that were affected. I provide a detailed timeline in Appendix Table B1. In the case of the Broadcom b43 driver, the project began early 2003, the first working driver is released in late 2005, with the full driver was included in the Linux kernel in June 17, 2006, and updated to be 802.11 compliant in January 24, 2008. In total, there were 4,252 reviews for 93 products that used any of these SOCs in my Amazon review. Of these, 14 products became compatible with open source software.

2.4 Data collection and variable definition

[38] For instance, Intel (https://openwrtsummit.files.wordpress.com/2017/09/17-10-intel_openwrt_summit_sponsor_talk_v03.pdf), Technicolor (https://openwrtsummit.files.wordpress.com/2017/11/openwrt-40-technicolor.pdf), Qualcomm Atheros (https://wiki.codeaurora.org/xwiki/bin/QSDK/), Mediatek (https://labs.mediatek.com/en/platform/linkit-smart-7688), and Marvell (https://github.com/MarvellEmbeddedProcessors) all provide versions of OpenWRT.

2.4.1 Data Collection

I collect a unique dataset of wireless routers' product characteristics, open source compatibility, and product market performance by merging data from three sources. The first source is wikidevi.com, a crowdsourced website for hardware data, and contains the product characteristics of wireless routers released between 1999-2017. The second source(s) are the compatible devices lists from three open source custom router firmware projects: OpenWRT, LEDE[39], and DD-WRT. The third and final source consists of reviews, sales rank, and price data between 1996-2014 from Amazon.com.

In the first stage, I merge Wikidevi with OpenWRT, LEDE, and DD-WRT using their respective device names. I manually check the names to ensure positive matches. Secondly, I obtain the Amazon Standard Identification Number (ASIN) of these devices from Wikiedvi. Finally, in the third stage I use the ASINs to obtain all reviews for products, as well as product metadata such as descriptions and package dimensions from Amazon (McAuley et al. 2015). Again, I manually match the device names with product names from the Amazon dataset to ensure the ASIN match quality. The final dataset consists of 184,013 reviews for 1,106 wireless routers and wireless embedded products. Of these, 284 devices are compatible with OSS while 822 are never compatible.

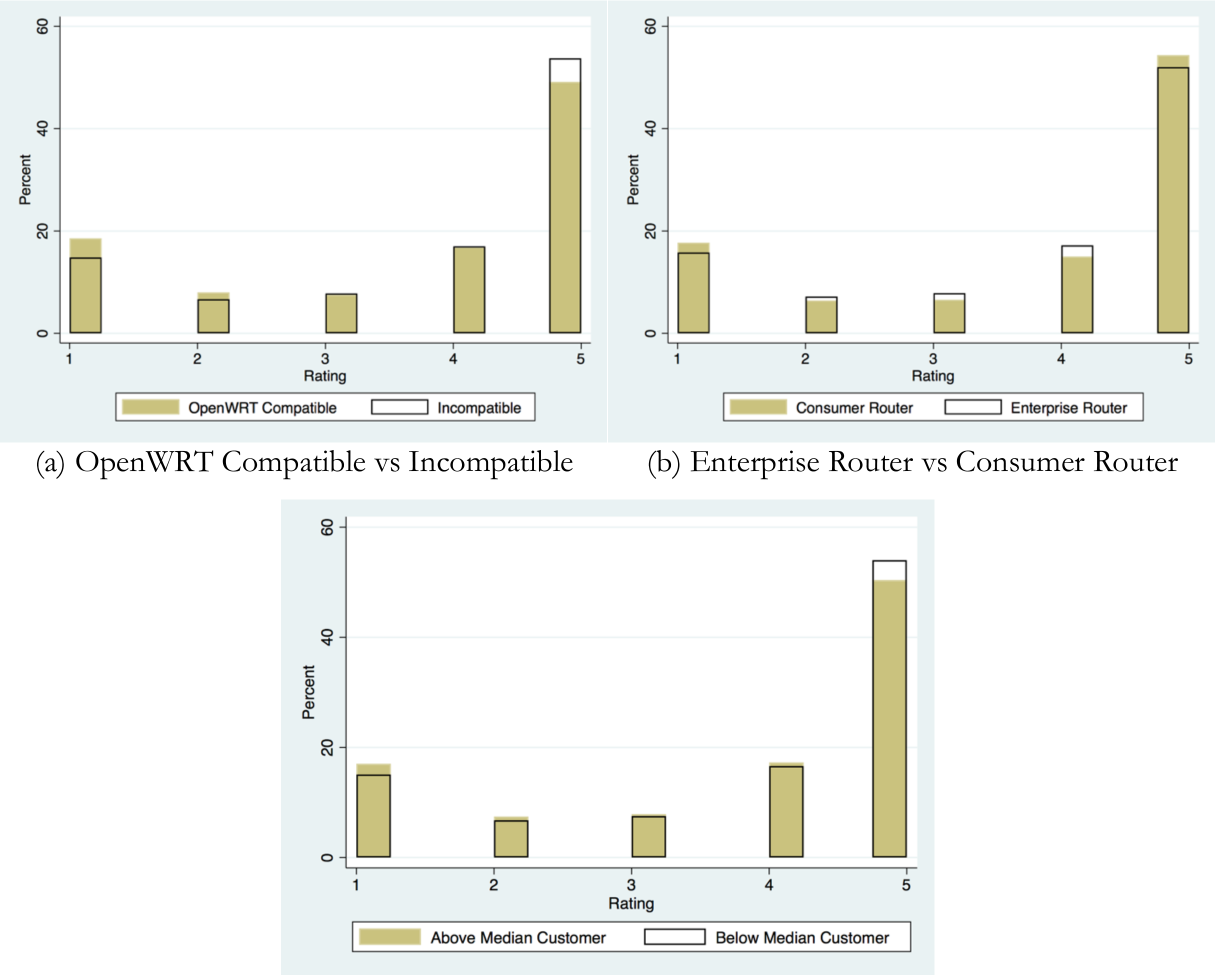
For most of the regressions, I am interested in a subset of routers that were exogenously made compatible with custom firmware. In these instances, I only use those routers that were released before the reverse engineering events. To further exclude firms' endogenous selection into compatibility, I compare treated product (product whose device drivers were reverse engineered) to a control group of products that were never treated. This brings my sample down to 217 devices and 12,787 reviews, with 72 devices whose device drivers were reverse engineered.

[39] As of January 2018, OpenWRT and LEDE have merged back to OpenWRT

2.4.2 Variables of interest

Dependent variables

From Amazon reviews, I obtain detailed data on users' perceptions of the value of a product. My main dependent variable is $Rating_{rit}$ is the star rating on review r for product i at time t. Ratings are integers ranging from 1-5. Figure 2.2 plots the distribution of review ratings for products that are compatible with the complementary open source software, and those that are never compatible.



(a) OpenWRT Compatible vs Incompatible (b) Enterprise Router vs Consumer Router

(c) Above median savvy consumer vs below median savvy consumer

Figure 2.2. Distribution of ratings for OSS compatible and proprietary products

Figure 2.2 plots the review ratings for OSS compatible versus proprietary products. Review ratings range from 1-5 stars. OpenWRT compatible routers are obtained from the compatible devices list of the OpenWRT project. Enterprise Routers are defined as products containing "employee," or "enterprise" in the product descriptions. Above Median Customer is defined as customers with above median purchase category cosine similarity.

I use three proxies for product market performance[40]. $ReviewCount_{it}$ is the total number of reviews for product i at time t. I define $MarketShare_{it}$ as the share of reviews for product i at time t, which I obtain by calculating $MarketShare_{it} = \frac{ReviewCount_{it}}{\sum_{i'} ReviewCount_{i't}}$. Finally, I rank each product according to the number of review counts they receive at time t to obtain $Ranking_{it}$. To code this variable, I use the "rank(), track" function[41] of the egen command in Stata which ranks the lowest value as rank 1.

Finally, I measure review informativeness from the number of users that found the review helpful. Amazon.com allows users give individual reviews a binary rating of either helpful or not. Furthermore, Amazon provides this information to its users, allowing users to sort reviews based on how helpful they are. Prior literature has shown how review helpfulness varies with other review characteristics (Mudambi and Schuff, 2010). I define $Informative_{rit}$ as the number of users that found review r helpful. Alternatively, in robustness checks I use the fraction of users that found the review to be helpful, as well as a log transformation of the number of users that found the review helpful.

OpenWRT Compatibility and Reverse Engineering

[40] Prior work using customer review data has shown the relationship between customer reviews and product sales (Forman paper, Hitt and Li, Chen and etc, Mayzlin and Chevalier, Lee and Hosanagar, Luca and Zervas,). Review data and sales are closely linked, with prior research finding that books that receive higher ratings lead to increased sales (Chevalier and Mayzlin 2006). However, competitive aspects of reviews may lead to fake reviews that may bias analyses using reviews (Luca and Zervas 2016).

[41] One benefit of using the track rank is that positive coefficients can be interpreted as improvements in performance.

The matched dataset between wireless routers and open source repositories allows me to identify products that are compatible with the various versions of open source firmware. I code OSS_i as a dummy variable indicating whether router i is listed on the compatible devices list of either DD-WRT, OpenWRT and LEDE. If compatible, I define $OSSDate_i$ as the date at which router i became compatible. Finally, at the review level, I define $PostOSS_{rit}$ as whether review r for product i was left at time $t \geq OSSDate_i$.

My main independent variable is an indicator variable denoting whether a router was exogenously made compatible with an open source operating system. I collect a list of 25 SOCs from 4 reverse engineering events. The list of reverse engineered driver names were obtained through an extensive online search of wireless drivers[42]. The list of compatible SOCs were obtained from matching the wireless driver names to SOC models listed the Linux Wireless Wiki[43]. I define $Treated_i$ as a dummy variable if router i uses a wireless SOC whose device drivers were reverse engineered.

Enterprise and consumer routers

I categorize routers into enterprise grade or consumer grade. For each router, I obtain the Amazon.com product description as well as the title of the product. I define $Enterprise_i = 1$ if the product description or title contains the words "enterprise," "business," or "employee.[44]" In my dataset, there are a total of 138 routers that are thus classified as enterprise routers.

Text analysis

[42] I combine historical versions of "Wireless LAN resources for Linux" by Jean Tourrilhes (https://hewlettpackard.github.io/wireless-tools/), the Wikipedia page "Comparison of open-source wireless drivers" (https://en.wikipedia.org/wiki/Comparison_of_open-source_wireless_drivers), and through interviews with experts in the field. Detailed documentation is available in the Appendix.

[43] The Linux Wireless Wiki is the source for most Linux wireless drivers and can be found at https://wireless.wiki.kernel.org/, with a historic archive is located at http://linuxwireless.sipsolutions.net/en/users/Documentation/.

[44] These words were obtained through manual inspection of the data.

I take advantage of the rich text data contained in the reviews. I define $MentionCheap_{rit}$ as an indicator variable for whether review r contains synonyms for "cheap," and similarly I define $MentionExpensive_{rit}$ as an indicator variable for whether review r contains synonyms for "expensive." I obtain a list of synonyms from Thesaurus.com[45]. Finally, I define $MentionOSS_{rit}$ as whether review r mentions OpenWRT, DD-WRT, ASUS WRT, or Tomato[46].

Finally, I preprocess review text and extract keywords using term frequency-inverse document frequency (tf-idf). Recent advancements in machine learning and data retrieval have significantly reduced costs of text processing, paving the way for their use in strategic management research (Menon et al. 2018, Teodorescu 2017). Using standard procedures for text analysis, I clean and stem the review text. The appendix provides detailed procedures for cleaning the text data. I obtain the tf-idf vector for all review text. Intuitively, the tf-idf captures idea that keywords are words that appear frequently in a given document but not in other documents. For those reviews that mention open source OpenWRT, I concatenate all the reviews and create a tf-idf vector for the concatenated document.

Reviewer data

In addition to changes in review text, I am interested in a compositional change of reviewers following an open source shock. In particular, I am interested in a measure of the technical sophistication of customers. Because I have detailed data on reviewers, I am able to identify reviewer characteristics through their review history. For each reviewer, I obtain a measure of the reviewer's tech "savviness" by comparing review history similarities with a select group of technically able users. I detail the process in further detail.

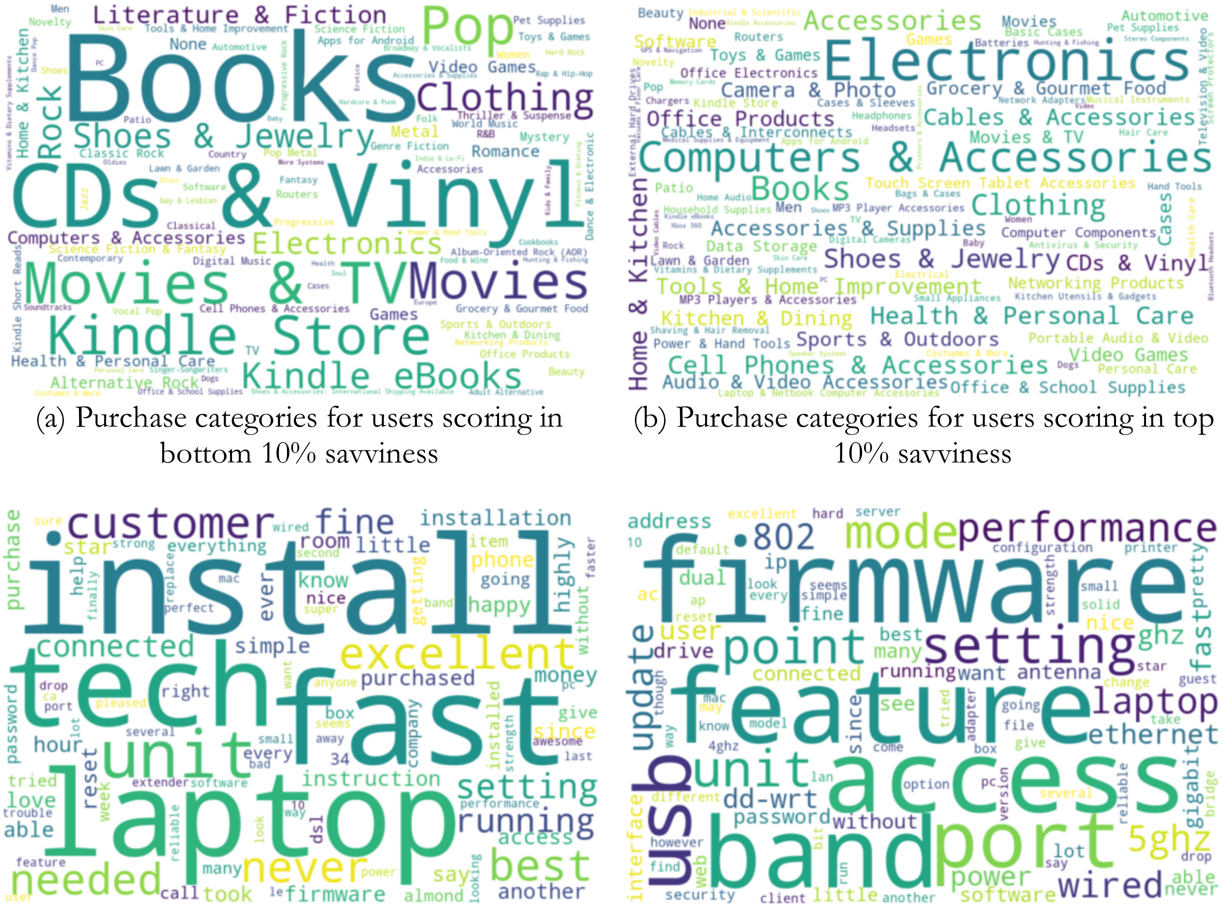
[45] For cheap, I use the following list of words: cheap, competitive, economical, low cost, low price, reasonable, bargain. For expensive, I use the following list of words: expensive, costly, extravagant, fancy, lavish, overpriced, pricey, upscale, and valuable.

[46] AsusWRT and Tomato are popular variants of the OpenWRT software.

I collect the review history of all reviewers who mention open source software (i.e., $MentionOSS_{rit} = 1$). This review history data is not limited to wireless routers, but consists of all reviews ever left by that reviewer on Amazon. There are a total of 3,821 reviews and 3,357 reviewers that meet this criterion. I create a purchase category history vector for each reviewer: Each row in the vector corresponds to a purchase category (e.g., Electronics, wireless router, etc.), and each corresponding element is the number of times a reviewer has left a review for a product in that category. I average their purchase history vectors to create a technologically "savvy" reference review history vector that tracks the purchasing reviewers.

For all 151,270 reviewers in my dataset, I measure the cosine similarity between the reference savvy vector and the reviewer's purchase history[47]. Cosine similarity is a measure often used in machine learning (Teodorescu 2017), and collaborative filtering (Linden et al. 2003) to calculate the similarities between two vectors. Figure 2.3 shows a word map of the topics that these reviewers are interested in.

[47] An individual's savviness vector is calculated via a two-step process: The first step can be thought of as "training" a model to find users that are likely to use custom firmware; the second step is the application of the model. First, I create a "reference savvy vector." Starting with the entire set of reviewers and their histories, I collect a subset of reviewers that left a review for wireless routers. Within this subset, I again subsample those reviewers that leave reviews mentioning the use of custom firmware. I will refer to this set of reviewers as the "baseline savvy users." Using the baseline savvy users, I average their review category histories: each reviewer's review history can be seen as a point on the vector-space of review categories. Their coordinates correspond to the number of reviews they leave in that category. For example, if reviewer A left 5 reviews for Electronics, 2 reviews for Books, and no other reviews, her review history will be a point p = (5,2,0,0,…) on the product category space. I take the average of all baseline savvy users' review histories to create $\boldsymbol{p}^*$, the "baseline savvy vector." In the next step, I calculate the cosine similarity between all reviewers' purchase categories $\boldsymbol{p_i}$ and the baseline savvy vector: $Savvy_i = \frac{p_i \cdot p^*}{\|p_i\| \|p^*\|}$



(a) Purchase categories for users scoring in bottom 10% savviness

(b) Purchase categories for users scoring in top 10% savviness

(c) tf-idf weighted review text for users scoring in bottom 10% savviness

(d) tf-idf weighted review text for users scoring in top 10% savviness

Figure 2.3. Word clouds to explain savviness measure

Figure 2.3 shows word clouds for purchase category vectors (a-b), and for tf-idf weighted review text (c-d). Savviness measure is defined as the cosine similarity between a given user's review history vector/tf-idf weighted review text vector and the average vector for users who mention custom firmware in their reviews.

2.5. Empirical specifications

2.5.1 Impact of complementary goods on product performance

Hypothesis 1 stated that openness leads to complement innovations, increasing perceived value. To test this, I examine whether custom firmware projects affected customer perception of product quality. Identification comes from the unanticipated releases of reverse engineered drivers and ensuing firmware compatibility. I estimate the following difference-in-differences model using ordinary least squares on the review data:

$$Rating_{rit} = \beta Post_{rit} \times Treated_i + \phi_i + \lambda_t + \epsilon_{rit} \qquad (1)$$

where $Rating_{rit}$ is the rating for review r for router i at time t. $Post_{rit}$ is an indicator for whether the review time t is left after router i is compatible with custom firmware. $Treated_i$ is an indicator for whether the device driver for router i was reverse engineered, making it exogenously compatible with the complementary good. In all specifications I include router fixed effects ϕ_i, and month fixed effects λ_t. The coefficient of interest, β captures the increase in review ratings caused by open source compatibility. It measures the difference in ratings for routers that were exogenously made compatible with OpenWRT, compared to routers that are never compatible.

In addition to the router and month fixed effects, I include other router-time-varying controls. The router and month fixed effects are powerful controls, accounting for changes in ratings by time-invariant characteristics of individual routers, as well as review trends within the Amazon platform. To control for router specific changes in ratings, I include linear and quadratic terms in the router age. Additionally, I include review-specific controls for whether the router contains mentions of price or open source.

2.5.2 Reviewer characteristics and information

Hypothesis 2 consists of two parts, first part of which states that savvy reviewers leave more informative reviews. To test whether savvy reviewers have more information, I model the relationship between a user's savviness and the characteristics of the reviews they leave. In particular, I am interested in how helpful reviews are to other customers. I estimate the following equation using ordinary least squares.

$$\boldsymbol{Informativeness_{rit} = \beta Savviness_{rit} + \gamma X_{rit} + \lambda_t + \phi_i + \epsilon_{rit}} \qquad (2)$$

The outcome $Informativeness_{rit}$ measures the informativeness of a review, for which I use two measures: Number of users that found the review helpful, and Review length. $Savviness_{rit}$ measures the cosine similarity between reviewer for review r and the reference "savvy" purchase category. In addition to fixed effects for calendar month and routers, I control for linear and quadratic terms for router age, and review text content. In addition to the informativeness, I also test whether savviness is related to the months since release, as well as the overall rating.

The main coefficient of interest β measures how review informativeness varies with the savviness of the reviewer. Again, controlling for the product and calendar-month fixed effects will control for any time-invariant characteristics of the product associated with informativeness, as well as platform-wide changes in reviewers that elicits more helpful reviews. Furthermore, including trends for product age aims to control for the trends in helpfulness of reviews over the router's lifetime. Thus, $\beta > 0$ will be consistent with savvy reviewers leaving more information. To the extent that review informativeness is driven by router-specific time trends, I include linear and quadratic terms for the router's age.

The second part of Hypothesis 2 states that complement innovations attract savvy users. I estimate whether open source compatibility caused a change in the customer composition for those products. These regressions will allow me to uncover the mechanisms underlying the changes in product market performance. Towards this I estimate the following equation using ordinary least squares.

$$\boldsymbol{Savviness_{rit} = \beta Post_{rit} \times Treated_i + \phi_i + \lambda_t + \epsilon_{rit}} \quad (3)$$

$Savviness_{rit}$ measures the cosine similarity between reviewer r and the reference savvy purchase category[48]. $Post_{rit}, Treated_i$ are defined as before. I include product fixed effects ϕ_i and month fixed effects λ_t. The coefficient of interest β measures a compositional change in the demand for product i following the availability of a downstream complementary good. A positive coefficient implies savvy customers become more likely to leave reviews for product i once it is open source compatible, whereas a null effect suggests there is no sorting of customers.

2.5.3 Strategic implications

Hypothesis 3 discusses the strategic implications for openness and complement innovation. Specifically, I ask whether openness has heterogeneous effects across different product types. I test for two relationships: First, heterogeneity in the impact of complement innovations on review helpfulness; second, heterogeneity in the impact of complement innovations on product sales. I explain the tests in more detail below.

I first explore how openness differentially affected the helpfulness of reviews. The statistical relationship I intend to uncover is whether savvy users' reviews mitigate informational asymmetries more for complex products (enterprise routers) than for less complex products. For this, I estimate the following regression specification:

[48] I abuse notation here since r is at the review level, not the reviewer level. However, 83.77% of reviewers leave only one review for wireless routers in my time period, and reviews and reviewers can be considered one-to-one.

$$Informative_{rit} = \beta Post_{rit} \times Treated_i \times Enterprise_{rit} + \gamma_2 Post_{it} \times Treated_i + \phi_i + \lambda_t + \epsilon_{rit} \qquad (4)$$

$Informative_{rit}$ is defined as above to be the number of users that found the review helpful. $Post_{rit}$ and $Treated_i$ are defined as above, and $Enterprise_{rit}$ is an indicator variable for whether router i is an enterprise router. I include router (ϕ_i) and calendar month (λ_t) fixed effects. Note that other first and second order interaction terms are subsumed by the product and time fixed effects as above.

Finally, I test whether savvy users' helpful reviews helped to enhance product sales. As stated previously, I proxy for sales using review based measures. Towards this, I estimate the following triple differences equation.

$$Performance_{rit} = \beta Post_{rit} \times Treated_i \times Enterprise_i + \gamma Post_{it} \times Treated_i + \phi_i + \lambda_t + \epsilon_{rit} \qquad (5)$$

$Post_{rit}, Treated_i$ are all defined as above. $Performance_{rit}$ is a measurement of product market performance, for which I use measures of review rank, count, and share. $Enterprise_i$ is a dummy variable for whether router i's description contains enterprise keywords. Note that because the control group consists of never-treated routers, none of the control group has post period observations. This implies $Post_{it} \times Enterprise_i$ is perfectly collinear with the triple interaction term, and therefore it is not included in the above specification. Similarly, indicators for $Post_{rit}, Treated_i$, $Treated_i \times Enterprise_i$ and $Enterprise_i$ are subsumed by the product and time fixed effects. The main coefficient of interest β captures the differential increase in ratings for enterprise routers compared to the consumer routers. γ captures the effect of the complementary good on non-enterprise routers.

2.6 Results

2.6.1 Changes in review ratings

I document changes in review ratings following open source compatibility. I compare the changes in ratings to the average ratings of never compatible routers, as specified in equation 1. The results are presented in Table 2.1. Column 1 plots the most basic specification, only controlling for product and month fixed effects. We see that open source compatibility increases router review ratings by 0.527 stars, consistent with our hypotheses. Controlling for a linear trend in the product age does not significantly alter the effect size (Column 2). In Columns 3-5, I include measures for the reviewers' savviness. Controlling for whether the reviews mention certain topics (cheap, expensive, and open source) do not affect the ratings, but have a positive effect on ratings once reviewer characteristics are controlled for.

Table 2.1. OSS increases user review ratings

	(1)	(2)	(3)	(4)	(5)
Dep. Var.			Review Rating		
Post x Treated	0.625^{**}	0.624^{**}	0.617^{**}	0.584^{**}	0.575^{**}
	(0.264)	(0.264)	(0.262)	(0.261)	(0.257)
Product age		-0.000	-0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)	(0.000)
Savviness			0.294^{**}		0.289^{**}
			(0.131)		(0.131)
Review text similarity			-0.554		-0.700^{*}
			(0.403)		(0.407)
Mention cheap				0.080	0.088
				(0.067)	(0.066)
Mention expensive				0.203^{***}	0.215^{***}
				(0.070)	(0.070)
Mention OSS				0.210	0.235
				(0.211)	(0.206)
Product FE	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y
Observations	12759	12759	12759	12757	12757
Adjusted R^2	0.141	0.141	0.142	0.141	0.142

Robust standard errors (clustered at the product-month level) in parentheses. Observations at the review level. Sample includes products affected by reverse engineered drivers and products never compatible with open source software. Post indicates the review was left when the product was compatible with open source software. Reverse engineered indicates the product used hardware whose drivers were reverse engineered. Product age denotes the product age in days. *Savviness* measures the cosine similarity between the reviewer's review history and the review history of reviewers that mention custom firmware in their reviews. *Review text similarity* measures the cosine similarity between the tf-idf vector of a review and the tf-idf vector of reviews who mention OSS. *Mentions cheap* is an indicator for whether the review contains synonyms of cheap (cheap, economical, competitive, reasonabl [sic], low price, low cost, bargain), while *Mentions expensive* is an indicator for whether the review contains synonyms of expensive (expensive, costly, extravagant, fancy, lavish, overpriced, pricey, upscale, valuable). *Mention OSS* is an indicator for whether the review mentions open source firmware project names (OpenWRT, DD-WRT, Tomato, AsusWRT). Sample sizes change in column (3) because cosine similarity is missing for 34 reviewers, review text is missing for 22 individuals. Note that variables for Post OSS and Reverse engineered drop out in the specification because they are collinear with Month fixed effects and Product fixed effects respectively.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To address concerns of parallel trends, I plot a dynamic coefficient plot for the diff-in-diff results. Figure 2.4 shows the difference in ratings between treated and control group routers across time, relative to the time they become compatible with custom firmware. The baseline difference between the two groups is taken one year before compatibility. I see that before custom firmware is compatible, there are no significant differences in review ratings between the treatment and control group routers, consistent with the assumption of parallel trends. Upon custom firmware compatibility, ratings increase for treated routers relative to custom routers.

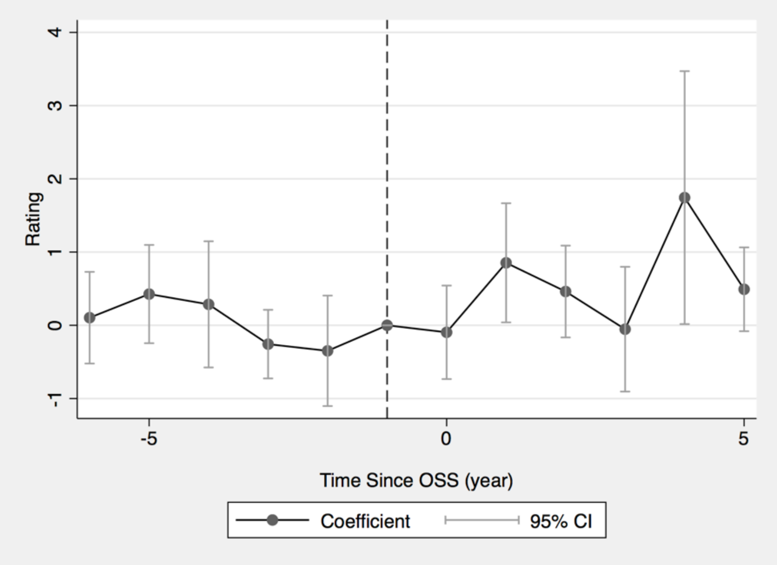


Figure 2.4. Trends for review ratings and review counts

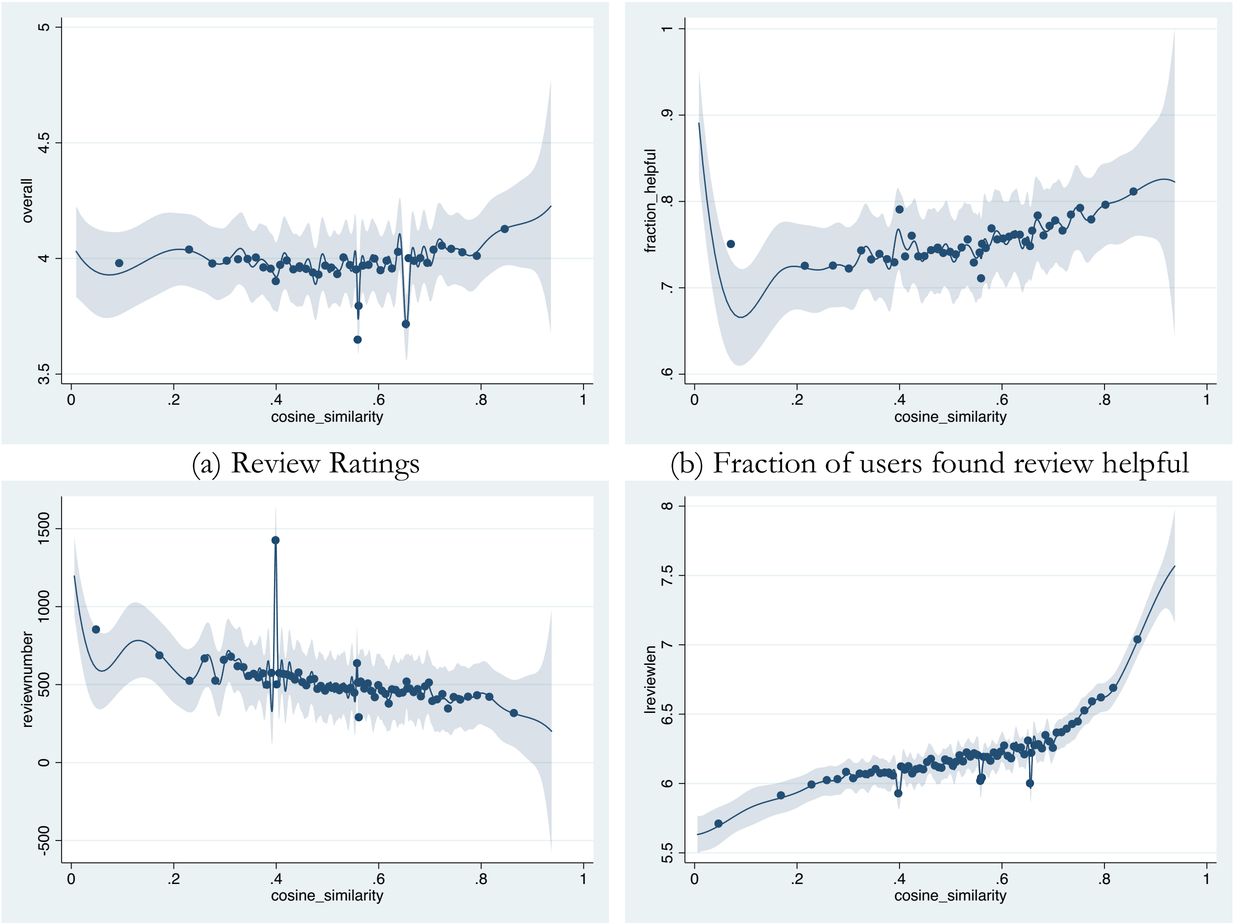
Figure 2.4 plots the dynamic coefficient graph for routers that were exogenously made compatible with open source firmware. Each point on the graph denotes the difference in ratings for exogenously open sourced routers and control group routers at each point in time.

2.6.2 Mechanisms for product performance change

Figure 2.4 plots binned scatterplots of the relationship between savviness and various review characteristics: Review rating, Fraction of users who found the review helpful, the order of the review (number of reviews that precede the current review), and log review length. All binned scatterplots control for router fixed effects, thus showing variation within a product[49]. Visually, reviewers' savviness seem uncorrelated with their review ratings (panel (a)), but are positively

[49] The plots were generated using the binsreg command in stata.

associated with review helpfulness (panel (b)), as well as review length (panel (d)). Savvy reviewers also leave reviews earlier in the product life cycle (panel (c)), suggesting they share many characteristics with lead users.



(a) Review Ratings

(b) Fraction of users found review helpful

(c) Number of reviews prior to focal review

(d) Log review length

Figure 2.5. Relationships between savviness and review characteristics.

Figure 2.5 plots binned scatterplots between reviewer savviness and review characteristics. The x-axis indicates savviness, or the cosine similarity of the reviewer's purchase categories with savvy users' purchase categories. Panel (a) plots the relationship between cosine similarity and review ratings; (b) with the fraction of reviews who found the review helpful; (c) with the timing of the reviews – earlier reviews have fewer reviews prior to the review; and (d) with log review character length. All binned scatterplots are generated using the binsreg command in Stata and include router fixed effects.

More formally, I test the visual relationships using ordinary least squares. Table 2.2 presents results from an OLS regression. Each specification includes router fixed effects as well as calendar-month fixed effects. Confirming the visual hypotheses, I observe that reviewer savviness is not significantly associated with review ratings (Column 1). On the other hand, more customers find savvy users' reviews helpful. A one standard deviation increase in reviewer savviness is associated with 2.77 more users finding the review helpful (Column 2). Similarly, a one standard deviation increase in savviness is associated with an 8% decrease in the review order (Column 3), as well as a 40% increase in the review length.

Table 2.2. Relationship between savviness and review characteristics

	(1)	(2)	(3)	(4)
	Rating	Fraction found helpful	Log(Review order)	Log(Review length)
Savviness	0.005	0.093***	-0.166***	0.860***
	(0.055)	(0.012)	(0.028)	(0.050)
Age (months)	-2146.654	1491.269	8591.162	-2173.731
	(17217032.614)	(29488032.987)	(41137478.126)	(6810167.868)
Age sq. (months)	0.000**	-0.000***	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Month FE	Y	Y	Y	Y
Product FE	Y	Y	Y	Y
Observations	183853	183853	183853	183853
Adjusted R^2	0.104	0.161	0.906	0.169

Standard errors in parentheses, standard errors clustered at the router-month level. Observations are at the review level. Savviness measures the cosine similarity between the purchase category history of the reviewer and the average purchase category of users who mention open source firmware in their reviews. Age measures the time in months since the first review for a router. Rating is the star rating left for a product in a review, and ranges from 0 to 5. Found helpful measures the number of Amazon users who found the review helpful. Review order is the ordering of the review for a given product, with the first review being 1, second review being 2, and so forth. Review length is the string length of the review.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In addition to documenting how review content varies with savviness, I show open source compatibility attracts savvy customers. Table 2.3 presents results from estimating equation 2. Columns 1-3 use "Savviness" as the dependent variable, while Columns 4-6 use indicator variables for whether the reviewer ranks above median savviness (Column 4), the top 25% (Column 5), and top 10% (Column 6). First, I find the average savviness of reviewers increases following open source compatibility. In all specifications, there is a positive and significant effect of open source compatibility on reviewer savviness. Open source compatible routers are 18.2 percentage points more likely to be reviewed by an above median savvy reviewer (Column 4), and 23.8 percentage points more likely to be reviewed by a top 25% savvy reviewer (Column 5). The effects at the extreme savvy customers is muted.

Table 2.3. OSS shifts customer base

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	Purchase category similarity			Above median	75th percentile	90th percentile
Post x Treated	0.064***	0.064***	0.053***	0.134**	0.175***	0.038
	(0.019)	(0.019)	(0.017)	(0.067)	(0.054)	(0.037)
Age		-0.000	-0.000	-0.000	-0.000***	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mentions cheap			0.015***	0.020	0.049**	0.067***
			(0.006)	(0.021)	(0.023)	(0.014)
Mentions expensive			0.005	0.006	0.025	0.055**
			(0.008)	(0.022)	(0.027)	(0.022)
Mentions OSS			0.060***	0.107	0.174***	0.129**
			(0.005)	(0.017)	(0.018)	(0.017)
Product FE	Y	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y	Y
Observations	12759	12759	12757	12757	12757	12757
Adjusted R^2	0.185	0.185	0.186	0.089	0.099	0.026

Robust standard errors in parentheses, clustered at the product-month level. Observations at the review level. Post denotes a dummy variable indicating whether the product was compatible with custom firmware at that time. Treated is an indicator for whether the product uses hardware whose drivers were reverse engineered.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.4 further shows that savvy users are indeed benefiting more from the complement innovations. Columns 1 and 2 show the result of estimating equation 1 on the savvy customer subset, and columns 3 and 4 show the result of estimating equation 1 on the non-savvy customer subset. While the coefficient on the interaction term is positive and significant for savvy users, it is not significantly associated with increased review ratings for non-savvy users. However, a t-test comparing the coefficient sizes cannot reject the null that the coefficients are the same magnitude.

Table 2.4. Perceived value depends on adoption costs of customers

	(1)	(2)	(3)	(4)
Dep. Var.	Review Rating			
Subsample of users:	Above median savviness		Below median savviness	
Post x Treated	0.640^{**}	0.608^{**}	0.238	0.268
	(0.299)	(0.284)	(0.277)	(0.276)
Savviness		1.504^{***}		-0.071
		(0.291)		(0.204)
Review text similarity		-0.790^{*}		-1.221
		(0.462)		(0.768)
Mention cheap		0.115		0.052
		(0.073)		(0.115)
Mention expensive		0.197^{**}		0.302^{***}
		(0.090)		(0.087)
Mention OSS		0.169		0.025
		(0.190)		(0.395)
Product FE	Y	Y	Y	Y
Month FE	Y	Y	Y	Y
Observations	7740	7738	4970	4970
Adjusted R^2	0.141	0.146	0.152	0.153

Robust standard errors in parentheses, clustered at the product level. Observations are at the review level. The dependent variable is the review star rating from Amazon. Columns (1-2) present results from a regression in the subsample of reviewers who have above median savviness, and reviewers with below median savviness in Columns (3-4).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.6.3 Strategic implications

Thus far, I have argued that opening IP and the ensuing complement innovations increase product market performance. Furthermore, complement innovations attract savvy users, who possess information that is helpful to other customers. This section explores the strategic implications for reductions in information imperfections. Specifically, I test whether the information imperfection reduction is greater for more complex products such as enterprise routers. I first test whether openness leads to more helpful reviews for enterprise products, and then show suggestive evidence of differential impacts on product market performance.

Table 2.5 presents a modified form of equation 4 by segmenting the reviews into four categories: 1) reviews for enterprise products by savvy reviewers, 2) reviews for enterprise products by non-savvy reviewers, 3) reviews for non-enterprise products by savvy reviewers, 4) reviews for non-enterprise products by non-savvy reviewers. Column 1 shows that other customers find the reviews left by savvy users on enterprise products to be more helpful. A one standard deviation increase in savviness is associated with 1.42 more customers finding the review helpful. On the other hand, customers do not find reviews for non-savvy customers helpful for enterprise products (Column 2), nor do they find reviews for non-enterprise products helpful (Columns 3-4). A t-test comparing the coefficient sizes for columns 1 and 2 is significant at the 1% level, suggesting that within savvy users, more information is shared for enterprise products.

Table 2.5. Stronger information effects for enterprise products

	(1)	(2)	(3)	(4)
	Enterprise		Non-Enterprise	
	Savvy	Non-Savvy	Savvy	Non-Savvy
Dep var: Found helpful				
Post x Treated	2.845^{**}	1.327	-0.215	2.601
	(1.326)	(2.278)	(1.417)	(3.112)
Age	528.161	-1899.228	958.244	-6173.709^{***}
	(18179338.190)	(44473764.521)	(36548004.055)	(0.000)
Age squared	0.002^{***}	0.001^{***}	0.001	0.001
	(0.000)	(0.000)	(0.001)	(0.034)
Mention cheap	2.039	0.470	3.585^{**}	0.442
	(1.337)	(1.890)	(1.510)	(0.901)
Mention expensive	5.328	0.965	4.003	3.150
	(4.537)	(0.759)	(2.564)	(2.898)
Mention OSS	-1.450	2.629	2.233	-0.190
	(1.658)	(2.056)	(1.608)	(0.893)
Observations	1806	960	5912	3984
Adjusted R^2	-0.006	0.144	0.037	0.047

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Finally, I test whether reductions in information imperfections can lead to product market performance. The results from estimating equation 5 appear in Table 2.6. Each column utilizes a different performance measure for companies: the raw number of reviews (1), the log of the number of reviews + 1 (2), the share of reviews (3) the rank of the firm (4), and the log rank of the firm (5). The main coefficient of interest is the triple interaction term which documents the differential benefit to enterprise products, above the effect to non-enterprise products. Complement innovations lead to increased reviews for enterprise products, compared to non-enterprise products (Columns 1-2). Raw counts may not be appropriate if the market is expanding, thus leading to an increase in overall number of reviews over time. The market share (column 3) and the rank within a given month (Columns 4-5) are less susceptible to changes in overall market trends. For all outcome variables, complement innovations increased the number of reviews (and by inference, sales,) more for enterprise products than for non-enterprise products.

Table 2.6. Benefits to complements are greater for enterprise products

	(1)	(2)	(3)	(4)	(5)
Dep. Var.	Review count	Log(1+Review count)	Review share	Ranking	Log(1+ Ranking)
Post x Treated x Enterprise	0.845***	0.345***	0.004***	14.863**	0.263**
	(0.136)	(0.096)	(0.001)	(7.207)	(0.103)
Post x Treated	-0.128	-0.210***	-0.000	-14.942***	-0.200***
	(0.269)	(0.057)	(0.001)	(3.091)	(0.031)
Product FE	Y	Y	Y	Y	Y
Month FE	Y	Y	Y	Y	Y
Observations	13306	13306	13306	13306	13306
Adjusted R^2	0.319	0.444	0.493	0.598	0.801

Robust standard errors in parentheses, clustered at the product level. Observations are at the product-month level. Post denotes a dummy variable indicating whether the product was compatible with custom firmware at that time. Treated indicates whether a product used a device driver that was reverse engineered, exogenously allowing for compatibility with custom firmware.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.7 Conclusion

This study examines how firm and product market performance were affected following the loss of intellectual property. I find the loss of IP allowed for the creation of a downstream complementary good, which ultimately benefited customers. Product market performance for products , with both customer ratings and sales increasing. At the firm level, however, small firms were positively affected by these changes, while large firms' performance decreased. This suggests that one mechanisms by which opening IP can benefit firms is through lowering adoption costs and driving increases in sales.

This study contributes to a growing literature on ecosystem management, and in particular, the importance of complementors in focal firm profitability. A growing literature looks at the impact of how the focal firm's incentives are affected by complementors (Adner and Kapoor 2010, Jacobides et al. 2018, Kapoor and Furr 2015). I contribute to this literature by illuminating a demand side mechanism by which focal firms can benefit from introducing complementary goods (Rietveld and Eggers 2018). Furthermore, I illuminate how complementary goods allow firms to unlock different types of complementarities, adding support to the literature uncovering the determinants of network effects (Afuah 2013, Lee et al. 2006, Suarez 2004). In particular, I combine findings from the information systems literature to add more nuance to the traditional debate of value creation and value capture in openness research.

This study also contributes to a large literature on the benefits of open source software. Specifically, I show that there exist heterogeneity in the benefit from open source. In addition to the supply side determinants of the benefits from open source (Nagle 2015, 2018), I document that demand side characteristics can drive the benefit from open source. While open source software on average leads to complementary innovations (Boudreau 2010, Wen et al. 2015, Zhang 2016), the benefits are realized by a smaller number of users. This suggests the need for a broader understanding of who benefits from open source software, and how to better measure its economic impacts (Greenstein and Nagle 2014).

Finally, I make a methodological contribution towards measuring demand heterogeneity. Most studies on demand heterogeneity do not have direct measures of consumer heterogeneity (Rietveld and Eggers 2018). Building on the recent adoption of machine learning techniques in the strategic management literature (Kaplan and Vakili 2015, Menon et al. 2018, Teodorescu 2017), I develop methods to measure demand side heterogeneity. In particular, by introducing techniques used in other fields such as collaborative filtering (Linden et al. 2003), as well as text analysis and natural language processing, I am able to quantify the amount of heterogeneity present in the demand side.

References

Adner R (2017) Ecosystem as structure: an actionable construct for strategy. *Journal of Management* 43(1):39–58.

Adner R, Kapoor R (2010) Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations. *Strategic management journal* 31(3):306–333.

Afuah A (2013) Are network effects really all about size? The role of structure and conduct. *Strategic Management Journal* 34(3):257–273.

Alexy O, West J, Klapper H, Reitzig M (2018) Surrendering control to gain advantage: Reconciling openness and the resource-based view of the firm. *Strategic Management Journal* 39(6):1704–1727.

Andersen E (2018) Microsoft joins Open Invention Network to help protect Linux and open source. Retrieved (November 8, 2018), https://azure.microsoft.com/en-us/blog/microsoft-joins-open-invention-network-to-help-protect-linux-and-open-source/.

Baldwin CY, Clark KB (2000) *Design rules: The power of modularity* (MIT press).

Boseley S (2010) Glaxo offers free access to potential malaria cures. *The Guardian* (January 20) https://www.theguardian.com/science/2010/jan/20/glaxo-malaria-drugs-public-domain.

Boudreau K (2010) Open platform strategies and innovation: Granting access vs. devolving control. *Management science* 56(10):1849–1872.

Brandenburger AM, Nalebuff BJ (2011) *Co-opetition* (Crown Business).

Bresnahan TF, Greenstein S (1999) Technological competition and the structure of the computer industry. *The Journal of Industrial Economics* 47(1):1–40.

Brodkin J (2016) FCC forces TP-Link to support open source firmware on routers. *Ars Technica.* Retrieved (September 30, 2018), https://arstechnica.com/information-technology/2016/08/fcc-forces-tp-link-to-support-open-source-firmware-on-routers/.

Butler D (2010) GlaxoSmithKline goes public with malaria data. *Nature.*

Casadesus-Masanell R, Llanes G (2011) Mixed source. *Management Science* 57(7):1212–1230.

Casadesus-Masanell R, Llanes G (2015) Investment Incentives in Open-Source and Proprietary Two-Sided Platforms. *Journal of Economics & Management Strategy* 24(2):306–324.

Cassia F (2006) Hackers put Linux back into Linksys WiFi routers | TheINQUIRER. *http://www.theinquirer.net.* Retrieved (September 18, 2018), https://www.theinquirer.net/inquirer/news/1015215/hackers-linux-linksys-wifi-routers.

Chan R (2018) Microsoft will let anyone use 60,000 of its key software patents as it moves to play more nicely with open source developers. *Business Insider.* Retrieved (October 15, 2018), https://www.businessinsider.com/microsoft-oin-patents-open-source-linux-2018-10.

Chevalier JA, Mayzlin D (2006) The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* 43(3):345–354.

Cohen WM, Levinthal DA (1990) Absorptive capacity: a new perspective on learning and innovation. *Administrative science quarterly*:128–152.

Corbet J (2005) LinkSys courts Linux hackers with WRT54GL (LinuxDevices) [LWN.net]. Retrieved (September 18, 2018), https://lwn.net/Articles/162429/.

Corbet J, Rubini A, Kroah-Hartman G (2005) *Linux Device Drivers: Where the Kernel Meets the Hardware* (O'Reilly Media, Inc.).

Evans DS, Hagiu A, Schmalensee R (2008) Invisible engines: how software platforms drive innovation and transform industries (MIT press).

Gawer A, Henderson R (2007) Platform owner entry and innovation in complementary markets: Evidence from Intel. *Journal of Economics & Management Strategy* 16(1):1–34.

Golson J (2014) Tesla Just Gave All Its Patents Away to Competitors. *Wired* (June 12) https://www.wired.com/2014/06/tesla-just-gave-all-its-patents-away-to-competitors/.

Greenstein S, Nagle F (2014) Digital dark matter and the economic contribution of Apache. *Research Policy* 43(4):623–631.

Henkel J (2006) Selective revealing in open innovation processes: The case of embedded Linux. *Research policy* 35(7):953–969.

Henkel J, Schöberl S, Alexy O (2014) The emergence of openness: How and why firms adopt selective revealing in open innovation. *Research Policy* 43(5):879–890.

Hu B, Hu M, Yang Y (2017) Open or Closed? Technology Sharing, Supplier Investment, and Competition. *M&SOM* 19(1):132–149.

Jacobides MG, Cennamo C, Gawer A (2018) Towards a theory of ecosystems. *Strategic Management Journal.*

Kaplan S, Vakili K (2015) The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal* 36(10):1435–1457.

Kapoor R, Agarwal S (2017) Sustaining superior performance in business ecosystems: Evidence from application software developers in the iOS and Android smartphone ecosystems. *Organization Science* 28(3):531–551.

Kapoor R, Furr NR (2015) Complementarities and competition: Unpacking the drivers of entrants' technology choices in the solar photovoltaic industry. *Strategic Management Journal* 36(3):416–436.

Katz ML, Shapiro C (1985) Network externalities, competition, and compatibility. *The American economic review* 75(3):424–440.

Kumar V, Gordon BR, Srinivasan K (2011) Competitive strategy for open source software. *Marketing Science* 30(6):1066–1078.

Lee E, Lee Jeho, Lee Jongseok (2006) Reconsideration of the winner-take-all hypothesis: Complex networks and local bias. *Management Science* 52(12):1838–1848.

Linden G, Smith B, York J (2003) Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* (1):76–80.

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* 62(12):3412–3427.

Macher JT, Mowery DC (2004) Vertical specialization and industry structure in high technology industries. *Business strategy over the industry lifecycle.* (Emerald Group Publishing Limited), 317–355.

Mannes J (2017) Facebook open sources Caffe2, its flexible deep learning framework of choice. *TechCrunch.*

McAuley J, Targett C, Shi Q, Van Den Hengel A (2015) Image-based recommendations on styles and substitutes. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* (ACM), 43–52.

Meeker H (2005) Open Source and the Legend of Linksys | News | LinuxInsider. Retrieved (September 8, 2018), https://www.linuxinsider.com/story/43996.html.

Menon A, Choi J, Tabakovic H (2018) What You Say Your Strategy Is and Why It Matters: Natural Language Processing of Unstructured Text. *Academy of Management Proceedings.* (Academy of Management Briarcliff Manor, NY 10510), 18319.

Metz C (2015a) Google Just Open Sourced the Artificial Intelligence Engine at the Heart of Its Online Empire. *Wired* (November 9) https://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/.

Metz C (2015b) Facebook Open Sources Its AI Hardware as It Races Google. *Wired* (December 10) https://www.wired.com/2015/12/facebook-open-source-ai-big-sur/.

Musk E (2014) All Our Patent Are Belong To You. Retrieved (September 24, 2018), https://www.tesla.com/BLOG/ALL-OUR-PATENT-ARE-BELONG-YOU.

Nagle F (2015) Crowdsourced Digital Goods and Firm Productivity: Evidence from Open Source Software. *Harvard Business School Research Paper* (15–062).

Nagle F (2018) Learning by Contributing: Gaining Competitive Advantage Through Contribution to Crowdsourced Public Goods. *Organization Science*.

Parker G, Van Alstyne M (2017) Innovation, openness, and platform control. *Management Science*.

Parker GG, Van Alstyne MW (2005) Two-sided network effects: A theory of information product design. *Management science* 51(10):1494–1504.

Pash A (2006) Turn your $60 router into a $600 router. *Lifehacker*. Retrieved (September 17, 2018), https://lifehacker.com/178132/hack-attack-turn-your-60-router-into-a-600-router.

Rietveld J, Eggers JP (2018) Demand Heterogeneity in Platform Markets: Implications for Complementors. *Organization Science* 29(2):304–322.

Rogers EM (2003) *Diffusion of Innovations* (Simon and Schuster).

Schreier M, Oberhauser S, Prügl R (2007) Lead users and the adoption and diffusion of new products: Insights from two extreme sports communities. *Marketing Letters* 18(1–2):15–30.

Suarez FF (2004) Battles for technological dominance: an integrative framework. *Research Policy* 33(2):271–286.

Suarez FF (2005) Network effects revisited: the role of strong ties in technology selection. *Academy of Management Journal* 48(4):710–720.

Teodorescu MH (2017) Machine Learning Methods for Strategy Research.

Toyota (2015) Toyota Opens the Door and Invites the Industry to the Hydrogen Future | Toyota USA Newsroom. Retrieved (November 8, 2018), http://corporatenews.pressroom.toyota.com/releases/toyota+fuel+cell+patents+ces+2015.htm.

Undercoffler D (2015) CES 2015: Toyota opens patents on hydrogen fuel cell technology - Los Angeles Times. *latimes.com*. Retrieved (November 8, 2018), http://www.latimes.com/business/autos/la-fi-hy-ces-2015-toyota-patent-hydrogen-fuel-cell-20150104-story.html.

Von Hippel E (1986) Lead users: a source of novel product concepts. *Management science* 32(7):791–805.

Wen W, Ceccagnoli M, Forman C (2015) Opening Up Intellectual Property Strategy: Implications for Open Source Software Entry by Start-up Firms. *Management Science*.

West J (2003) How open is open enough?: Melding proprietary and open source platform strategies. *Research policy* 32(7):1259–1285.

Zhang L (2016) Intellectual property strategy and the long tail: Evidence from the recorded music industry. *Management Science* 64(1):24–42.

3. Do Corporate Contributions Spur Innovative Activity in Online Crowdsourced Goods? Evidence from the Linux Kernel.

Do Yoon Kim, Mike Teodorescu

Abstract

A significant part of the digital economy is built upon "digital public goods". Despite significant advances in our understanding of the factors that affect corporate provision of digital public goods, the converse question of how corporate participation affects innovation in digital public goods is less understood. In this paper, we test how increased corporate participation affects source code contribution and follow-on innovation in a large open source project. We test whether corporate contributions increase non-corporate contributions or whether they "crowd out" non-corporate contributions. We utilize a unique data set of contribution behavior from a large open source project to test our hypotheses. We find that while corporate contributions attract more contributions (both from other corporations and non-corporate entities), they also lead to less follow-on innovation. Using a shock that increased corporate commercialization incentives, we find that increased corporate contributions have a "crowding out" effect on individual contributions, specifically to corporate authored general-purpose code. These results suggest that increased corporate participation may limit innovations in online digital contexts.

3.1 Introduction

A significant part of the digital economy is built upon crowdsourced digital goods. Apache, an open source software that powers many servers, is estimated to account for up to $12 billion in mis-measured productivity (Greenstein and Nagle 2014). The Linux kernel "runs 90 percent of the public cloud workload, has 62 percent of the embedded market share, and 99 percent of the supercomputer market share" (Corbet and Kroah-Hartman 2017). Wikipedia, a crowdsourced encyclopedia, is often ranked as one of the top 10 most visited websites in the world (Zhang and Zhu 2011). The ubiquity of crowdsourced goods has attracted governmental attention as well, with the European Union announcing its plan to adopt more open source software (Leroux 2017).

As crowdsourced digital goods become more popular, corporate stakes in such goods have also increased. The Linux Foundation reports that well over 85 percent of all kernel development is demonstrably done by developers who are being paid for their work (Corbet and Kroah-Hartman 2017). Such "free" contributions are often made by corporations, despite the potential of competitors' free-riding. Management scholars have tackled this question using several lenses, from selective revealing (Ahuja et al. 2013, Henkel 2006, Henkel et al. 2014), cost savings (Gambardella and von Hippel 2019), standard setting (Simcoe 2012), and organizational learning (Nagle 2018a). The broader platforms and ecosystems literature also suggests firms benefit from increased adoption of their platforms, subsequently allowing them to capture value (Boudreau 2010, Parker and Van Alstyne 2005, West 2003).

Despite significant advances, little light has been shed on how corporate participation affects the direction of innovation in crowdsourced digital goods. In this paper, we shed light on the converse question of how corporate participation affects innovation in crowdsourced digital goods. We build on the literature on innovation incentives in public (e.g., academic) and private sectors (Aghion et al. 2008, Murray 2010, Stern 2004). Studies highlight the importance of considering the dynamics between content and community (Kane and Ransbotham 2016, Nagaraj 2017), but a comprehensive picture of how public and private sectors can affect dynamics is sparse (Huang and Murray 2009).

We point out the potentially divergent effects of corporate participation. On the one hand, increased corporate participation can attract increased contributions[50] from individuals who are motivated by career concerns, potentially shifting the overall direction of innovation in the corporations' favor. On the other hand, existing motivations to for individuals to contribute may be crowded-out by such developments, limiting the overall level of development. In either circumstance, corporate innovations themselves will be driven by immediate commercialization incentives, and thus lead to less follow-on innovation and more exploitative behavior[51].

We test our hypotheses on a detailed panel dataset of file level changes to the source code in a large open source project. Using the entire commit history of the Linux kernel from 1991-2018, we create a monthly panel dataset of changes made to each file in the kernel, for 103,110 files and 2.6 million observations. Furthermore, because we observe the committer's emails, we are able to discern whether the individual is affiliated with a corporation, or if they are non-corporate entities[52]. Finally, we measure follow-on innovations by mapping the function call structures into dependency graphs (MacCormack et al. 2006), and various source code quality metrics.

Our first set of correlational evidence suggests corporate contributions attract more contributions from both corporate and non-corporate entities but lead to less follow-on innovation. This pattern is driven by several factors. First, non-corporate contributions generally lead to more

[50] I define contributions as either "commits" to a file within a git repository, or the number of lines added to that file. A commit is a unit of discrete change in the source code.

[51] I define follow-on innovation to a focal file as the number of other files that utilize the focal file. Various files can utilize the focal file by, for instance, reusing (importing) routines contained within the focal file.

[52] We proxy direct ties between corporations and individuals through their email addresses. Details of variable definition and the potential limitations are discussed in section 4.

follow-on innovation. While corporate created files are able to attract non-corporate contributions toward hardware-specific source code, corporations seem to “crowd-out” non-corporate contributions in general purpose source code. Second, corporate files display greater self-contributing behavior, which is generally associated with lower levels of follow-on innovation.

Simple correlations between measures of innovation and corporate/non-corporate contributions to the Linux kernel may be biased by the overall demand for Linux, as well as the underlying “quality” of the idea being implemented. We utilize a sudden increase in corporate investment into the Linux kernel in November 2007 to identify the causal impact of increased corporate participation. We argue that files created by corporations in this time period will directly embody the increasing commercialization incentives compared to non-corporate files created during the same time period.

We contribute to the literature on innovation in crowdsourced digital goods. In particular, we contribute to a growing literature documenting how crowdsourced digital goods evolve (Kane and Ransbotham 2016, Nagaraj 2017). We uncover an additional source of limitations to crowdsourced innovations that originates in different incentives facing individuals and corporations. This suggests the potential necessity of government participation in maintaining the exploratory nature of crowdsourced digital goods. We also contribute to the literature on the benefits from contributing to, and using open source software (Lerner and Tirole 2002, Nagle 2018a, b). Finally, we contribute to the literature on innovations and private/public streams of knowledge (Aghion et al. 2008, Murray and Stern 2007, Stern 2004).

3.2 Literature Review and Theory

3.2.1 Literature Review

Initial interest in crowdsourced goods focused on the individual incentives to exert considerable effort to develop a public good. A significant portion of studies highlight the importance of "scratching an itch" (Von Hippel and Von Krogh 2003). On the other hand, economists advanced the argument for career concerns as the main driver of individual participation in public goods (Lerner and Tirole 2002). The literature has expanded significantly, uncovering mechanisms that limit crowdsourced content. Particularly relevant to our setting, content may itself serve as a regulator for contributions (Kane and Ransbotham 2016), and ownership issues can crowd out contributions by individuals (Nagaraj 2017).

Recent work addresses the increasing number of corporate contributions to public crowdsourced goods. The literature highlights three distinct motivations to contribute to open source software. First, companies "selectively reveal" their innovations to capture value from the other parts of their value chain (Casadesus-Masanell and Llanes 2011, Gambardella and von Hippel 2018, Henkel 2006, Lerner and Schankerman 2010, West and Gallagher 2006). By selectively revealing, companies are able to obtain development support from the community (Henkel 2006), decrease the costs of inputs (Gambardella and von Hippel 2018), or profit from proprietary technologies that the company owns (Casadesus-Masanell and Llanes 2011, Llanes and de Elejalde 2013). Second, companies may benefit from contributions to public goods because of organizational spillovers from employee learning (Kumar et al. 2011, Nagle 2018a). Finally, firms may contribute to crowdsourced digital goods if this allows them to implement certain standards within their industry. Prior work has shown that increases in private interest in technologies can cause inefficient (Simcoe 2012).

A broader literature exists on innovation motives in the public and private streams of knowledge. Public streams of knowledge, or knowledge "disclosed into the public commons" (Huang and Murray 2009) encompass basic scientific research, and allow for follow-on use by virtually anybody. Private streams of knowledge, on the other hand, are protected by formal intellectual property protections such as patents. The motivations to contribute to public streams of knowledge can be starkly different from the motivation to contribute to private streams of knowledge, with individuals receiving nonmonetary benefits for contributions to the public stream (Roach and Sauermann 2010, Stern 2004). Theoretical research has focused on the optimal timing of the different streams of research, suggesting that corporate "focus" makes them more suited towards the commercialization stage of an innovation, and "ownership" makes basic science research more suited towards the initial stages of an innovation (Aghion et al. 2008). Empirical work has tested this hypothesis, uncovering how intellectual property protection can lead to decreased follow-on innovation (Aghion et al. 2010, Murray et al. 2016, Murray and Stern 2007).

3.2.2 Theory and Hypotheses

In this section, we outline hypotheses on how individual contributions will respond to increased corporate contributions. A wide variety of individuals with diverse sets of motivation participate in crowdsourced digital goods (Von Hippel 1986). Individuals furthermore sort into projects that fit their own motivations (Belenzon and Schankerman 2015), and increases in corporate participation will attract individuals with greater career concerns, leading to increased innovation. At the intensive margin, however, two counteracting forces exist. On the one hand, increased corporate involvement may shift individual contributions to more closely align the objectives of the firm (Baker 2000). Combined with the individual sorting effect, corporate involvement will lead to increased contributions that are aligned with corporate motives. This brings us to our first hypothesis:

H1(a): Increased corporate participation in crowdsourced digital goods will attract new contributions by non-corporate individuals, and their contributions will be more closely aligned with corporate contributions.

On the other hand, individuals who receive nonmonetary benefits by contributing to public goods may find their benefits crowded out from the increased level of activity (Stern 2004). The Linux kernel has its roots in the free software movement, and the inherent distrust of corporate motivations may impose ideological costs of fixing bugs (Raymond 1999). Furthermore, recent work suggests how information itself may act as a mechanism to moderate future contributions. First, increased corporate participation may lead to decreased ownership, reducing individuals' utility from making contributions (Nagaraj 2017). Second, as corporate (and total) contributions increase, there may be increased returns to simple consumption of the crowdsourced good, reducing its production (Kane and Ransbotham 2016). These bring us to an alternate hypothesis.

H1(a): Increased corporate participation in crowdsourced digital goods will reduce new contributions by non-corporate individuals, and their contributions will not be aligned with corporate contributions.

We next elaborate on follow-on innovation. A key distinguishing point between corporate (private) and community (public) innovations is the freedom to pursue personal agendas (Aghion et al. 2008). While community members are free to make contributions they see fit, corporate contributors will face constraints imposed by their employers. One consequence of the different incentives is that corporate research leads to less follow on innovation (Aghion et al. 2010, Murray et al. 2016, Murray and Stern 2007). In contrast, contributions by non-corporate sponsored individuals are further away from commercialization, thus may garner more follow-on usage. A corollary of this hypothesis is the generality of the knowledge created. Because corporate focus is directed towards research projects that are closer to commercialization, they will be more likely to be contributing to hardware specific portions. This brings us to our second hypothesis:

H2: Corporate contributions will on average lead to less follow-on innovation. Furthermore, corporate contributions will be more hardware specific than non-corporate sponsored contributions.

3.3 Background and Setting:

3.3.1 The Linux Kernel Project

The main setting for our study is the Linux kernel project. The Linux kernel started in 1991 when Linus Torvalds was a university student. Attracting contributions from a large number of developers, the Linux kernel quickly grew and became one of the largest open source software projects in the world. As of 2019, the Linux kernel contains 875,104 commits by over 20,000 unique individuals.

Corporate interest in Linux is significant. Belying the widespread notion of altruistic "hackers", arguably more than 85% of the contributions to the Linux kernel come from individuals who are being paid for their work (Corbet and Kroah-Hartman 2017). In addition to for-profit organizations, nonprofit consortia have formed around supporting the Linux kernel and the surrounding ecosystem. Such consortia boast billion dollars of shared value, and hundreds of member organizations, as well as individual supporters[53].

The kernel development process is hierarchical, and contributions are vetted by editorial staff, much as in journal publications. The process involves individual contributors submitting patches to the Linux kernel mailing list. Upon receipt, kernel maintainers (appointed by Linus

[53] For example, the Linux Foundation (https://www.linuxfoundation.org/about/) and Purpl (https://prplfoundation.org/current-members/) boast their membership.

Torvalds and others) suggest changes to the submitted source code. After the source code is vetted, it is "committed" to the kernel mainline. Nagle (2018) provides a detailed overview of the submission process of the Linux kernel.

3.3.2 Android and the Open Handset Alliance

The Android operating system refers to a particular set of software that, in conjunction, drive countless smartphones and embedded systems. In our setting, an operating system consists of a kernel and other associated programs[54]. The Android operating system was developed by Andy Rubin in 2003. In 2004, Google acquired Android, unveiling it in 2007, and the first commercial smartphones using Android began appearing in September 2008. Android is currently the best-selling operating system in the world

We focus on events surrounding the release of the Android operating system. Android's adoption was not the result of Google's standalone efforts, but rather a coordinated act across many suppliers. In 2007, the Open Handset Alliance was established to support Android and develop an open standard for mobile devices. Prior to its founding, most operating systems for embedded devices used proprietary operating systems (e.g., Nokia's Symbian OS, Wind River system's VxWorks) or in-house versions.

We argue that the introduction and eventual success of the Android operating system significantly affected corporate incentives for a subset of the Linux kernel. Specifically, it led to increased contributions to those parts of the kernel that ensure proper hardware-specific interactions with mobile phones and embedded devices. In contrast, the demand for developing general features

[54] This is the difference between an Android operating system (Linux kernel + bionic C library) versus a standard GNU/Linux operating system (Linux kernel + GNU C Compiler).

of the kernel is less. We exploit the variation in incentives to identify the effects of increased corporate participation, as we explain in detail below.

3.4. Data and methods

3.4.1 Data Collection

We collect detailed source code data from the Linux kernel git repository, a large open source project. The rich data allows us to observe changes to the source code at two levels of analysis: commits and files. I first explain broad characteristics of git, and the repository, then move on to detailed data collection process.

Git is a popular version control system that allows developers to collaboratively work on software projects, as well as track which developers made what kinds of changes to the source code. Git repositories contain the entire development history of a given project. A commit is a discrete change to the source code base, and each commit may involve changes in multiple files. For example, in the Linux kernel git repository, there are 875,612 commits between September 1991 and January 2019. Contained within each of those commits are changes to 101,646 files.

The data collection was conducted as follows. We obtain all commits to the Linux kernel between 1991 and 2018[55]. We collect the hash (unique identifier for commits), commit details (date, author, committer) via the "git log" command. Using the commit hashes, we obtain further details about the commit using the "git log –numstat" command. This command lists all the files that were modified, the number of lines inserted and deleted for a given commit. Furthermore, git differentiates simple movements of the files, so we are able to differentiate between file creations and file relocations or renames.

[55] The commit data prior to 2005 (when the kernel began using git) was appended onto the recent git repository using the git graft command. More details available through the authors.

In addition to the commit logs for the Linux kernel, we obtain detailed code analysis through a commercial program to create a panel dataset of dependencies at the file level[56]. For each month between 1991-2018, we revert the source code to the historical version corresponding to the first date of that month using the "git checkout" command. On each historical version, we run a dependency analysis, or what the literature refers to as a "call graph extractor" (Murphy et al. 1998). For a given source code, the program returns a network of dependencies between files. A dependency from file A to file B exists if file A "includes" or "imports" functions from file B[57]. The final resulting panel dataset thus records for any file i at time t, the number of other files that referenced file i, and the number of files that i references.

We also manually collect data on the authors' affiliation through their email address domains. In particular, the Linux Foundation keeps track of the corporate contributors to the Linux Kernel[58]. We use this as a starting point, and additionally check each of the email addresses and categorize them into corporate or non-corporate emails. In light on the importance of the incentives facing individuals working for corporations and (those that are not), it is crucial to check whether these individuals are actually not employed at the times. To the extent that there are individuals using their personal mail accounts when committing on behalf of their company, we will underestimate corporate participation in the Linux Kernel[59].

Finally, we aggregate the panel dataset to the file level. The dataset consists of 109,326 unique files, across 324 months, for a total of 2,588,184 observations. For each file-level time series of variables of interest, we aggregate to the file level and record 1) the initial value (e.g., for code-to-

[56] We use Scitools Understand as in MacCormack, Rusnack, and Baldwin (2006)

[57] Details can be found in p. 110 of https://scitools.com/documents/manuals/pdf/understand.pdf

[58] The gitdm tool contains a mapping of known corporations: git://git.lwn.net/gitdm.git

[59] In future work, we plan to randomly sample users and check their work histories at the time these commits were made.

comment ratios), 2) the final value (e.g., for the file directory), 3) the maximum value (e.g, for the number of files that reference the focal file). We also omit directories that contain fewer than 15 files, and directories such as "DOCUMENTATION" that are irrelevant to our current setting. The final dataset is thus a file-level dataset with 101,646 observations.

3.4.2 Independent variables

Corporate vs Non-Corporate. At the file level, we code whether the initial author is an individual based on the email addresses they provide. For each file, we collect the email address of the initial contributor. We manually check each email address and categorize them into corporations and non-corporations. We assume that if an author uses their corporate email address when contributing to the kernel, they are being compensated for their efforts by the corporation. For example, gmail addresses are categorized as a non-corporate emails, whereas "Samsung.com" would be categorized under corporate emails.

To better align our empirical measures with our theoretical constructs, we categorize not-for-profit organizations and communities of developers as non-corporate entities. This includes many large contributing organizations such as OSDL (Open Source Development Labs). Doing so allows us to capture the extent to which corporations and "individuals" have differing motivations. Incorrect categorization of individuals as corporations is more likely to the extent that corporate contributors may choose to use their personal addresses. In contrast, corporate email addresses are not available to non-corporate individuals. Such miscategorization will act to decrease this difference in motivations, thus leading us to underestimate our results.

Hardware specific vs general purpose. We code whether a file is general purpose or whether it is specific to the hardware. For each file, we record the last known file path (i.e., after all renaming and relocations), then collect the parent directory of the file path[60]. The Linux kernel is structured in a certain way that makes the categorization simple. In particular, the directories /drivers/ and /arch/ correspond specifically to source code that is specific to a piece of hardware, and specific to a particular CPU architecture. In contrast, directories such as /mm/ and /kernel/ correspond to memory management and core kernel scheduling components[61]. We categorize source code as general purpose if it resides within a directory that is not hardware specific, and hardware specific otherwise[62].

3.4.3 Dependent variables

We collect two broad categories of dependent variables for a given file. The first, contributions, measures development activity on the focal file. The second, follow-on innovation, measures development activity outside of the focal file that utilizes the focal file's functions. We distinguish between the two, as the former measures "incremental" innovations in that they follow the dominant design of the file structure (Anderson and Tushman 1986). The latter, on the other hand, measures "recombination" in that they are finding new ways to apply what is already known (Henderson and Clark 1990; Fleming 2001).

Contributions. For each file, we obtain the number of commits made to a file in a given month as well as the number of lines added to/deleted from the file[63]. Collectively, contributions refer to either commit counts, line additions/deletions, or code deltas. We specify exactly which type of contribution when appropriate.

[60] The parent directory is the first folder name in the file path. For instance, given a file /arch/m68k/install.sh, we would record the parent directory as /arch/.

[61] The directories "arch" and "mm" stand for "architecture dependent" and "memory management" respectively.

[62] We use the following directories to define hardware specific files: arch, drivers, firmware, fs, samples, tools.

[63] In version control systems (VCS) such as git, additions to the source code are made in discrete chunks called a "commit." Each commit contains many line additions/deletions to many different files.

We distinguish between contributions by corporations and non-corporate entities. In particular, to calculate the number of contributions by corporations, we add the number contributions by email addresses associated with corporate domains. The number of contributions by non-corporate entities, on the other hand, is calculated by subtracting this number from the total number of contributions to the file.

Finally, we collect information about self-contributions. Self-citations are well documented in the patent innovation literature (Alcacer and Gittelman 2006). Relatedly, corporations can self-contribute to its own files. For each file, we tag any contributions that are made by the same company[64] and add them to generate the number of commits/lines of code that are contributed by the initial authoring company. We discuss the issue of self-follow-on innovations in the next section.

Follow-on innovation. To measure follow-on innovation, we obtain "references" to files[65]. Contributors are free to use and combine existing files/functions to create new files. We collect cases in which the focal file is depended upon by other files to measure of follow-on innovation[66]. We collect monthly dependency analyses for each file, thus measuring how file dependencies change over time. At the file level, we aggregate the dependencies to show 1) the maximum number of follow-on innovation, and 2) the average number of follow-on innovations.

Self-referencing is also possible in the software context. For instance, in order to create complex programs that span multiple files, it is necessary to create multiple interlocking files. Such interconnections may occur within the same commit, or they may happen dynamically. Because of

[64] Note we are using the same company, not the same email address. We match based on the corporate domain.

[65] For a focal file, references to it indicate follow-on innovation, whereas references to other files (from the focal file) indicate knowledge reuse.

[66] Specifically, we utilize the dependency analysis from the Scitools Understand program which "indicate a build-time dependency but don't always indicate a logical dependency. (e.g., unused include files)."

the computational burden of re-analyzing the source code after each commit, we collect the number of self-follow-on innovations that happen in the same month as the initial file commit.

Other variables. When possible, we collect other information that pertains to the source code quality. We collect the comment-to-code ratio, a common metric of how well documented the code is. Comment-to-code is calculated for each file 1) at the time the file was created, and 2) when the file was last observed. We calculate two measures of complexity. The first is McCabe Cyclomatic complexity (McCabe 1976), which counts the number of decision points (e.g., for, while, etc.) and the number of cases, adding one for each. The second is Nesting complexity, which measures the maximum nesting level of control constructs. High complexity measures can help predict software defects, for instance, security vulnerabilities (Shin and Williams 2008)[67].

3.5 Empirical specifications

3.5.1 Contributions and Corporate Authorship

To test whether increased corporate participation affects non-corporate contributions, we analyze the data at the file level. We bin the files into files by corporate/non-corporate, and files for hardware-specific/general purposes. Our goal is to compare the rate of contributions to the files within each bucket. Towards this, we estimate the following regression specification using a negative binomial specification on two subsets of the data (hardware specific files and generic files)[68].

$$Contributions_i = f(\alpha + \beta Corporate_i + X_i + \epsilon_i) \quad (1)$$

[67] Shin, Yonghee, and Laurie Williams. "An empirical model to predict security vulnerabilities using code complexity metrics." *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*. ACM, 2008.

[68] Alternatively, we estimate a form with the interaction term to test whether the coefficient sizes are significantly different across the two samples.

Here, $Contributions_i$ measures the total number of commits for file i over its observable period. We further separate contributions into those made by corporate and non-corporate entities. $Corporate_i$ is an indicator variable denoting whether file i was initially created by a corporate entity. We use two subsets of the data, defined by a variable $Specific_i$ which indicates whether file i is located in a directory for hardware specific files.

The coefficient β thus measures how contributions to file i would change by switching from a non-corporate created file to a corporate one. For control variables, we include fixed effects for the month in which the file was created (Month FE), as well as fixed effects for the top level directory structure (Directory FE). Month fixed effects aim to absorb changes in contributions caused by time-specific trends within the Linux Kernel, and Directory fixed effects absorb the mean levels of contributions within each directory[69].

Since we observe all contributing email addresses, we are able to see differential effects of corporate authorship across corporate and non-corporate contributors. In all cases, we estimate separate impacts of corporate authorship on contributions by corporations and non-corporate entities ($CorpContributions_i$ and $NonCorpContributions_i$ respectively). The coefficient β then measures the association between corporate authorship and contributions by corporate entities and non-corporate entities respectively.

3.5.2 Follow-on Innovation and Corporate Authorship

[69] In negative binomial regressions, the interpretation of fixed effects is different from that of in linear regressions (Wooldridge 2010). We replicate our results using ordinary least squares, as well as Fixed-effects Poisson models.

Next, to test how follow-on contributions differ for corporate initiated and non-corporate initiated files (Hypothesis 2), we estimate a similar functional form, but with an alternative dependent variable:

$$FollowonInnovation_i = f(\alpha + \beta Corporate_i + X_i + \epsilon_i) \quad (2)$$

Again, $CorporateInit_i$ is an indicator variable denoting whether file i was initially created by a corporate entity. Again, we subset the data based on indicators for whether file i is located in a directory for hardware specific files. We include Month fixed effects as well as top-level directory fixed effects.

The coefficient β then measures the difference in the likelihood of having follow-on innovation (as defined as use of file i in other programs) when the file is authored by a corporate entity. We also estimate the specification for files that are $Specific_i$ to a certain hardware, and files that are non-specific.

3.5.3 First steps to identifying the effect of corporate contributions

Thus far, we have compared the file-level differences between corporate and non-corporate authored files. While mean differences may suggest that corporate participation impacts contributions and follow-on innovation, it is unclear what drives the differences. In particular, we may observe file level differences even in the absence of corporate authorship effects if, for instance, the chain of command in the Linux kernel has more stringent policies for files created by corporate email addresses.

To address these issues and get closer to mechanisms outlined in the hypotheses, we use the emergence of the open handset alliance as an event that increased corporate commercialization incentives. Our empirical strategy in this section is to compare file-level differences before and after the shock. In particular, we assume that files created by corporations will reflect the change in commercialization incentives, whereas files created by non-corporate entities will be unaffected. Thus, we use the characteristics of non-corporate created files as a baseline against which we judge corporate created files.

Our empirical strategy thus consists of a difference-in-differences approach that compares contributions to files with differing levels of career concerns. Our treatment group are files created by corporations, and our control group are files created by non-corporate entities. Our treatment period are months after the announcement of the Open Handset Alliance. Intuitively, career concern related contributions should have increased more for corporate files in the post-period than in the pre-period. We thus estimate the following regression specification:

$$Contributions_i = f(\beta_0 Corporate_i + \beta_1 Post_i + \beta_2 Corporate_i \times Post_i)$$

where $Contributions_i$ is the number of commits that i receives over its lifetime, $Corporate_i$ is an indicator for whether file i was created by corporations, and $Post_i$ is an indicator for whether i was created after the announcement of the Open Handset Alliance. We consider both contributions by corporations and non-corporate entities.

The regression framework above compares the number of contributions to files created before and after OHA emerged. If significant career concerns exist, we should see an increase in the number of contributions by individuals, especially to corporate authored files ($\beta_2 > 0$). On the other hand, if corporate participation “crowds out” individual level contributions, we should see non-corporate contributions decrease ($\beta_2 < 0$). The coefficient β_0 captures the time-invariant difference in contributions between files created by corporations and non-corporate entities, and the coefficient β_1 captures the changes in contributions due to time.

As above, we estimate the regression using different subsets of the data: files created in hardware specific directories and hardware agnostic directories. Contributions to files in hardware specific directories requires understanding details about a specific bit of hardware that is usually corporation specific and cannot be reused. This suggests that in the presence of career concerns, contributions to hardware-specific files will increase, whereas in the presence of crowding-out effects, contributions to hardware-specific files will remain unchanged, whereas contributions by individuals to hardware agnostic files will decrease significantly.

Because the dependent variable is integer valued and highly skewed, we use a negative binomial to estimate the regression specification. When estimating interaction terms using nonlinear specifications such as the negative binomial, the coefficients may be biased (Athey and imbens 2006). Despite such drawbacks, prior work shows that nonlinear models for difference-in-difference estimates will have the same direction (Puhani 2012).

It is important to discuss the limitations of this approach. First, to the extent that there are non-parallel trends between corporate and non-corporate authored files, the effects will be confounded with the effects of increased corporate contributions. Second, there may be other events occurring simultaneously that confound the results.

3.6. Summary statistics and results

3.6.1 Broad trends for corporate and non-corporate source code

First, we show how the number of new files increased over time. Figure 3.1 shows the number of new files in a given year between 2001-2017. We see from the top blue line that the total number of new file creations is increasing over time. There is a noticeable peak between the years 2008-2009, but generally new file creation has been increasing linearly.

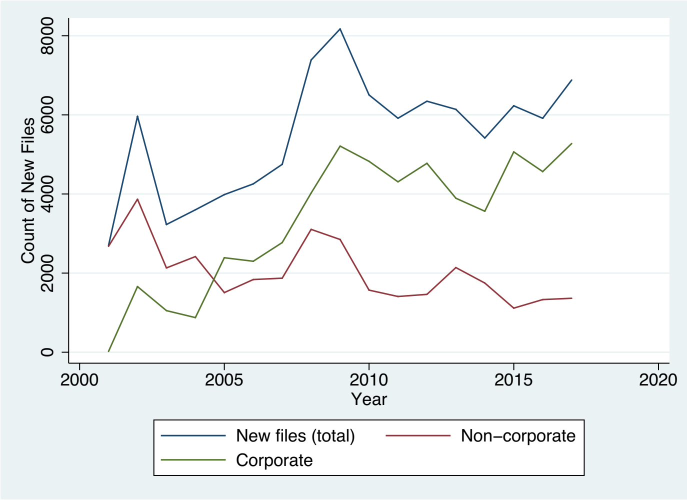
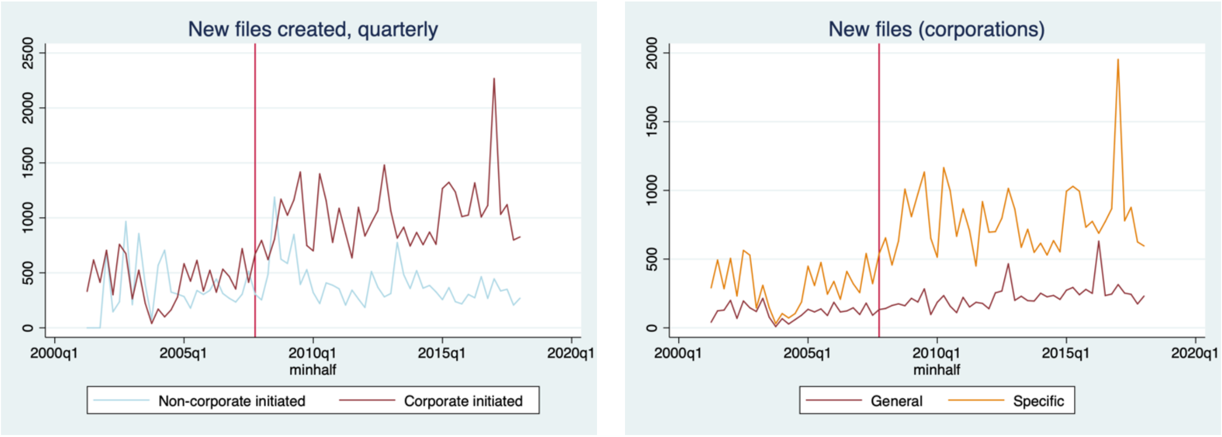


Figure 3.1. New file creation counts for the Linux Kernel

The lower two lines show that contributions by corporations are driving this increase, whereas new file creation by non-corporations has been on the decline. Between 2004-2005, the number of new files created by corporations surpassed the number of new files created by non-corporate entities. In 2017, over 80% of new file creations were by corporate entities, with 5,516 files out of 6,880 new file creation being driven by corporations. To the extent that corporate employees use non-corporate email addresses, this number underestimates the true extent of corporate participation in the Linux kernel.

Consistent with the emergence of the OHA being an increase in corporate commercialization incentives, we see corporate contribution levels increasing after the OHA is announced. Furthermore, we see in panel (b) of figure 3.2 that this increase is largely driven by hardware specific files.



(a) New files created overall

(b) Within corporate created files, are they hardware-specific or general purpose

Figure 3.2. New files created by corporate and non-corporate entities (a), and patterns of contribution within corporate authored files (b).

Below, we present summary statistics at the file level. We compare files created by corporations (column 1) and files created by non-corporate entities (column 2). Mean differences are shown in column 3 along with t-statistics. Overall, we find evidence consistent with corporate files embodying effort at the intensive margin: while there are fewer commits, there are more insertions and deletions to the code base. Corporations have a lower percentage of self-commits, likely because many non-corporate entities share common email hosts (e.g., gmail.com). Corporate created files attract further corporate activity, as the fraction of corporate commits is significantly higher.

Table 3.1. Summary Statistics and t-Test Results

	(1)		(2)		(3)	
	Corporate Initiated		Non-Corporate Initiated		Diff	
	mean	sd	mean	sd	b	t
Total commits	178.63	268.98	200.76	331.26	-22.14	(-11.33)
Total insertions	11737.63	34076.21	7645.11	21063.36	4092.52	(23.17)
Total deletions	5761.15	19754.96	5508.63	16966.63	252.53	(2.16)
Fraction self-commits	0.09	0.13	0.12	0.20	-0.03	(-25.85)
Fraction corporate-commits	0.53	0.13	0.50	0.17	0.03	(31.44)
Follow-on files	17.79	166.72	38.90	307.00	-21.12	(-9.48)
Follow-on entities	95.21	771.69	168.61	1378.17	-73.40	(-7.30)
Follow-on weight	289.52	1740.84	429.85	3363.38	-140.34	(-5.80)
Is specific	0.85	0.36	0.80	0.40	0.05	(19.41)
Comment-code ratio (initial)	1.25	16.20	1.24	10.43	0.01	(0.08)
Cyclomatic complexity (initial)	12.70	108.21	11.78	69.52	0.93	(1.41)
Nesting complexity (initial)	1.60	1.89	1.60	1.97	-0.00	(-0.11)
Comment-code ratio (final)	1.24	16.63	1.22	10.34	0.02	(0.21)
Cyclomatic complexity (final)	14.13	142.41	12.86	135.81	1.27	(1.23)
Nesting complexity (final)	1.62	1.92	1.62	2.01	-0.00	(-0.18)
Observations	55138		43807		98945	

Measures of follow-on innovation suggest that corporate files attract less follow-on innovation. The average corporate authored file has fewer follow-on files, entities, and weight. The average non-corporate created file is used by more than twice as many files as the average corporate created file. Relatedly, corporate files are more likely to be in directories where hardware specific code resides. Overall, the summary statistics suggest that corporations create files that are directed towards specific purposes that have little purpose elsewhere in the code base.

Corporate and non-corporate source code seem to have little difference in terms of code quality or complexity, both initially and eventually. The ratio of lines of comments to lines of code is similar for both files, suggesting that the code is equally well documented, both initially and eventually. We utilize various complexity metrics to measure how error prone the source code is. We see both McCabe cyclomatic complexity[70] and Nesting complexity are similar across both groups of files, suggesting no significant difference between how intricate the files are.

3.6.2 Changes to corporate and non-corporate source code over time

Next, we provide summary statistics to compare the 3 year window before and after the announcement of the open handset alliance.

[70] We use the maximum McCabe complexity across all entities within a given file. High McCabe complexity may indicate the code is prone to failure.

Table 3.2. Summary Statistics and t-Test Results (before and after 2008)

	(1)		(2)		(3)	
	Corporate Initiated		Non-corporate Initiated		Diff	
Before OHA (2005-2007)	mean	sd	mean	sd	b	t
Total commits	282.93	383.72	245.75	310.94	-37.18	(-5.97)
Total insertions	11471.14	34831.40	9160.92	25971.20	-2310.22	(-4.25)
Total deletions	8487.34	26970.65	6718.95	23428.23	-1768.39	(-3.90)
Fraction self-commits	0.06	0.11	0.05	0.07	-0.01	(-6.34)
Fraction corporate-commits	0.59	0.11	0.55	0.12	-0.05	(-23.02)
Follow-on files	31.98	216.83	36.41	302.63	4.43	(0.69)
Follow-on entities	151.10	796.33	151.78	1243.72	0.68	(0.03)
Follow-on weight	402.32	1839.69	401.65	2997.67	-0.67	(-0.01)
Is specific	0.81	0.39	0.79	0.41	-0.02	(-3.11)
Comment-code ratio (initial)	0.96	3.63	1.09	3.69	0.13	(1.71)
Cyclomatic complexity (initial)	11.76	68.33	11.69	107.94	-0.07	(-0.04)
# Lines (initial)	1.52	1.88	1.51	1.82	-0.01	(-0.37)
Comment-code ratio (final)	0.95	3.78	1.06	3.57	0.11	(1.51)
Cyclomatic complexity (final)	12.42	69.98	11.89	110.95	-0.53	(-0.27)
# Lines (final)	1.57	1.95	1.55	1.89	-0.02	(-0.46)
Observations	7287		5207		12494	
After OHA (2008-2010)						
Total commits	240.77	288.35	203.07	262.39	-37.70	(-9.64)
Total insertions	17582.14	40187.53	11574.75	29120.76	-6007.39	(-12.50)
Total deletions	9940.34	26104.06	6006.47	18201.54	-3933.87	(-12.84)
Fraction self-commits	0.07	0.11	0.06	0.09	-0.00	(-2.13)
Fraction corporate-commits	0.58	0.10	0.54	0.11	-0.04	(-26.08)
Follow-on files	15.60	147.83	23.38	214.60	7.78	(2.13)
Follow-on entities	104.01	939.51	110.80	1034.53	6.80	(0.36)
Follow-on weight	318.52	2140.68	339.68	2699.37	21.16	(0.45)
Is specific	0.86	0.34	0.87	0.33	0.01	(2.44)
Comment-code ratio (initial)	1.70	27.17	1.35	5.05	-0.35	(-1.33)
Cyclomatic complexity (initial)	14.07	122.15	8.64	22.82	-5.43	(-4.64)
# Lines (initial)	1.67	1.98	1.37	1.81	-0.30	(-9.95)
Comment-code ratio (final)	1.79	28.13	1.31	4.25	-0.48	(-1.80)
Cyclomatic complexity (final)	16.03	168.93	11.01	190.96	-5.01	(-1.70)
# Lines (final)	1.67	2.01	1.35	1.83	-0.32	(-10.58)
Observations	13900		7393		21293	

Comparing the pre and post tables, it seems that the OHA had the following effects on corporate created files. First, it increased effort at the intensive margin for corporate source code. Both insertions and deletions at the file level seem to have increased significantly, while the number of commits has stayed constant. Second, corporate source code has become significantly more targeted toward hardware specific goals. We see that follow-on innovation decreased significantly after the announcement of the OHA. However, while corporations have increased their creation of hardware specific source code, individuals have also contributed more hardware specific code. Finally, we see that corporate code has grown to become more well documented (more comments per line of code), and more complex (greater cyclomatic and nesting complexity).

3.6.3. Corporate files, hardware specificity, and self-contributions

We now formally test whether corporate initiated files attract more developmental activity, and if so, for which types of files. Towards this, we proceed as in section 4, splitting the sample into hardware specific or non-specific, then estimating the regression equations from before.

Table 3.3 shows results from estimating equation (1) using a negative binomial specification. We see that corporate authorship is associated with more contributions overall for both hardware specific and generic files (Columns 1 & 4). Distinguishing between contributions by corporations and non-corporations sheds some light on the observed difference between contributions to corporate and non-corporate files. Columns 2 & 5 present results using the contributions by other corporate entities as the dependent variable, while columns 3 & 6 present results using contributions by non-corporate entities as the dependent variable. We see that corporate files attract more contributions by corporations for both hardware specific and non-specific files. However, there is no corporate effect for non-corporate contributions in generic files.

Table 3.3. Corporate authorship and contributions

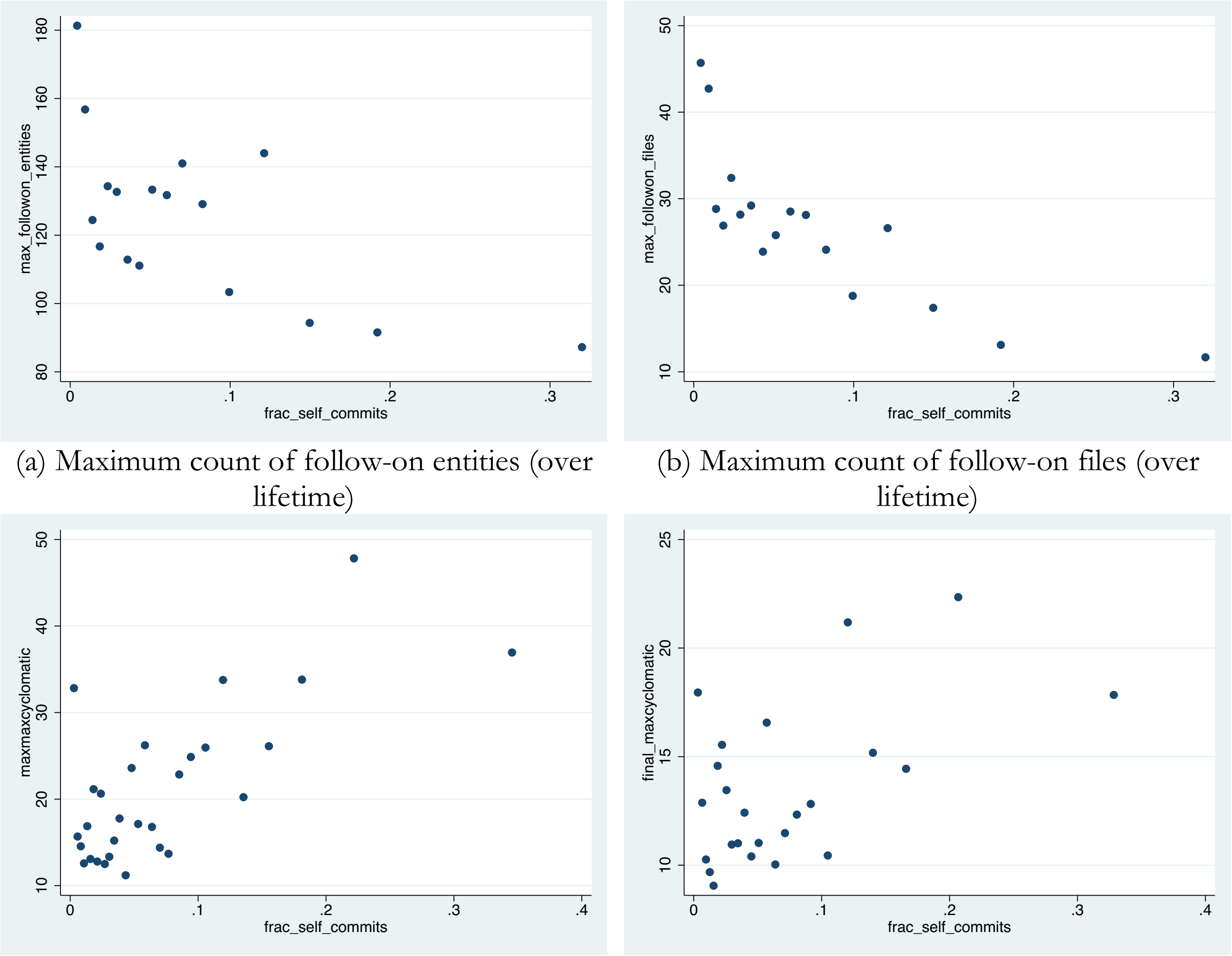
	(1)	(2)	(3)	(4)	(5)	(6)
	Hardware Specific Files			Generic Files		
Dep. Var: Contributions by:	Everyone	Corporate	Non-Corporate	Everyone	Corporate	Non-Corporate
CorporateInit	0.118***	0.217***	0.036***	0.094***	0.178***	0.018
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.477)
Constant	6.618***	5.929***	5.924***	3.286***	7.218***	3.263***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ln(alpha)	-0.100***	-0.076***	-0.033***	0.087***	0.156***	0.140***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Directory FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	81573	81573	81573	17340	17340	17340
Log-likelihood	-492159	-438208	-433633	-102356	-90711	-89531

p-values in parentheses, robust standard errors in parentheses. Dependent variables are the count of contributions (Everyone), count of contributions by corporate entities (Corporate), and non-corporate entities (Non-Corporate). CorporateInit is an indicator for whether the file is created by a corporation. Ln(alpha) tests whether the dispersion factor is significant from zero. We estimate two sets of regressions: for the subsample of files that are hardware specific, and the subsample of files that are not.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To further test for mechanisms, we will estimate equation (1) again, using self-contributions as the main dependent variable. Depending on the motivations for corporations, self-contributions should differ. If corporations are aiming to receive contributions for their products, we should see decreased self-contributions. On the other hand, if corporations are pushing forth an agenda, we should see increasing self-contributions.

Before moving into further details, we present stylized facts about self-contributions. Figure 3.3 shows binned scatterplots of the fraction of self-commits and various file characteristics. There is a noticeable negative correlation between the fraction of self-commits and follow-on innovation:

the more self-commits a file receives, the less likely it is to be used in other parts of the kernel. Similarly, self-commits are positively correlated with increases in code complexity. Taken together, these may indicate several points. First, self-contributions may facilitate the creation of complicated pieces of code that are useful for only a small subset of users. Second, from a development point of view, self-contributions may lead to overly complicated bits of code that are difficult for other entities to use, enhance, and further develop. We test these in more detail below.



(a) Maximum count of follow-on entities (over lifetime)

(b) Maximum count of follow-on files (over lifetime)

(c) Maximum cyclomatic complexity (over lifetime)

(d) Maximum cyclomatic complexity (last observed)

Figure 3.3. Relationship between fraction of self-commits and follow-on innovation (panels a & b) and source code complexity (panels c & d)

Table 3.4 presents results. We see that for hardware specific files, corporate files have a lower fraction of self-commits, but a greater number of self-commits overall. This suggests that for hardware specific files, while corporations are increasing self-commits to their files, the total number of commits are increasing faster than self-commits. Those non-self-commits seem to come both from corporate and non-corporate entities (Table 3.3, columns 2 & 3), but mainly from other corporations.

Table 3.4. Corporate files and self-contributions

	(1)	(2)	(3)	(4)
	Hardware Specific		Non-hardware specific	
	Fraction self-commits	Self-commits	Fraction self-commits	Self-commits
main				
Corporate	-0.006***	0.237***	-0.003*	0.219***
	(0.001)	(0.015)	(0.002)	(0.031)
Constant		4.001***		3.036***
		(0.131)		(0.481)
lnalpha		0.188***		0.053***
		(0.006)		(0.013)
Month FE	Yes	Yes	Yes	Yes
Directory FE	Yes	Yes	Yes	Yes
Observations	81570	81573	17333	17340
ll	44078.77	-271512.9	7461.93	-53549.03
Adjusted R	.201		.367	

Robust standard errors in parentheses. Coefficient estimates from estimating ordinary least squares (Columns 1 & 3) and negative binomial specifications (Columns 2 & 4). Corporate is an indicator for whether the file was initially created by a corporation. Sample sizes are smaller for columns 1 & 3 because the fraction of self-commits is not defined if there are zero commits.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In contrast, while general purpose corporate files also have more self-commits than their non-corporate counterparts, the total number of contributions does not seem to be increasing as quickly. On average, the fraction of self-commits is decreasing, but the effect is noisy. As we will later observe, corporate created files for a general purpose seem to be unable to attract contributions by non-corporate users, leading to decreased follow-on usage.

3.6.4 Corporate files and follow-on usage

Next, we document differences in follow-on usage for corporate and non-corporate files. Table 3.5 shows results from estimating equation (2). Column 1 shows that for hardware specific files, there is no statistical difference in follow-on innovation for corporate and non-corporate authored files. Thus, regardless of whether a file was authored by a corporation or a non-corporate entity, the likelihood of a hardware specific file being used in other parts of the Linux kernel are low. However, for generic files that are non-hardware specific, we see that corporate authorship is associated with significantly less follow-on innovation. In contrast to non-corporate entity authored files, corporate created files in general directories are used less than would be expected.

Table 3.5. Corporate files attract fewer follow-on innovation

	(1)	(2)
	Hardware Specific Files	Generic Files
Dep. Var:	Followon Entities	Followon Entities
CorporateInit	0.050	-0.385^{***}
	(0.395)	(0.000)
Constant	5.675^{***}	7.893^{***}
	(0.000)	(0.000)
lnalpha	1.052^{***}	0.969^{***}
	(0.000)	(0.000)
Month FE	Yes	Yes
Directory FE	Yes	Yes
Observations	35184	9037
Adjusted R^2		

p-values in parentheses. Negative binomial regressions. The number of observations is less than the number of total files, as ma ny files have no follow-on files.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Taken together with the prior results, corporate created files have different purposes depending on whether they are in hardware specific directories or generic directories. On the one hand, corporate authored hardware specific files receive significant investment from individuals, and are equally likely to lead to follow-on innovation as non-corporate created files. On the other hand, corporate authored generic files do not receive as much contributions by non-corporate entities, and do not lead to follow-on innovation as their non-corporate authored files. Broadly, corporate authored files are more likely to attract a larger number of commits, both self-commits and corporate commits.

3.6.5 Effects of increased corporate participation

In this section, we take a first step towards causality. We estimate the impact of increased corporate participation on the incentives to innovate by individuals and other corporations by comparing the file characteristics before and after the emergence of the OHA. We take a revealed preferences approach and document the observed contributing behavior in the presence of corporate activity.

Table 3.6 shows the results from estimating equation (2). The coefficient term in column (1) suggests there is no significant difference between the total number of commits for corporate and non-corporate files, before and after OHA emerged. Thus, increased corporate participation has not led to significant changes in contribution behavior. The main effects for time are negative and significant, suggesting a general decrease in contributions over time. There is a positive coefficient on the corporate term, suggesting that corporations generally attract more contributions.

Table 3.6. Corporate files and contributions

	(1) Total	(2) Individual	(3) Corporate
main			
Post x Corporate	-0.025	-0.054**	0.040*
	(0.025)	(0.026)	(0.025)
Post	-4.130***	-3.610***	-4.903***
	(0.136)	(0.153)	(0.274)
Corporate	0.136***	0.072***	0.189***
	(0.023)	(0.024)	(0.022)
Constant	4.920***	4.103***	4.335***
	(0.072)	(0.075)	(0.069)
lnalpha	-0.264***	-0.178***	-0.276***
	(0.005)	(0.005)	(0.006)
Month FE	Yes	Yes	Yes
Directory FE	Yes	Yes	Yes
Observations	73824	73824	73824
ll	-435455.540	-385212.888	-384469.501

p-values in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Columns 2 and 3 show how increased corporate participation led to differing incentives for other corporations and individuals. The interaction term in column 2 is negative and significant, suggesting that increased corporate participation led to a decrease in individual contributions to corporate created files. In contrast, column 3 shows there is a slightly positive effect on corporate contributions. In aggregate, the increased corporate participation has crowded out individual contributions.

We further divide the sample into specific and generic file types to measure which types of files are leading to decreases in the number of contributions. Columns 1-3 in Table 3.7 show the effect of increased corporate participation on general files, while columns 4-6 show the impact on hardware specific files. Interestingly, we see most of the negative effect on individual contributions is being driven by general files (column 2). In all other columns, the interaction terms are statistically insignificant, and do not allow us to reject a null effect.

Table 3.7. Corporate files and contributions

	(1)	(2)	(3)	(4)	(5)	(6)
		General			Specific	
	Total	Individual	Corporate	Total	Individual	Corporate
main						
Post x	-0.120*	-0.182***	-0.029	-0.005	-0.023	0.051*
Corporate						
	(0.064)	(0.066)	(0.063)	(0.027)	(0.029)	(0.027)
Post	-4.008***	-3.569***	-4.552***	-4.056***	-3.552***	-4.813***
	(0.242)	(0.254)	(0.244)	(0.129)	(0.147)	(0.270)
Corporate	0.155***	0.117*	0.182***	0.130***	0.057**	0.190***
	(0.059)	(0.061)	(0.057)	(0.025)	(0.026)	(0.024)
Constant	6.246***	5.446***	5.664***	4.825***	4.019***	4.226***
	(0.249)	(0.257)	(0.240)	(0.062)	(0.066)	(0.058)
/						
lnalpha	-0.295***	-0.190***	-0.319***	-0.272***	-0.190***	-0.282***
	(0.014)	(0.014)	(0.014)	(0.006)	(0.006)	(0.006)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Directory FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10894	10894	10894	62930	62930	62930
ll	-	-	-	-	-	-
	63334.237	55953.303	55638.866	371600.320	328708.051	328359.291

p-values in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.7 Conclusion and discussion

In this paper, we shed light on the converse question of how corporate participation affects innovation in crowdsourced digital goods. We build on the literature on innovation incentives in public (e.g., academic) and private sectors (Aghion et al. 2008, Murray 2010, Stern 2004). Studies highlight the importance of considering the dynamics between content and community (Kane and Ransbotham 2016, Nagaraj 2017), but a comprehensive picture of how public and private sectors can affect dynamics is sparse (Huang and Murray 2009). We argue that increased corporate participation will attract new individuals who are motivated by career concerns, and will additionally shift existing individuals to contribute innovations more aligned with corporate interests. Furthermore, we argue that these innovations will be more myopic, and thus lead to less follow-on innovation, shifting the crowdsourced innovations to become more exploitative.

We first provide correlational evidence of differences between corporate and non-corporate authored files. Overall, corporate files have more contributions that are directed towards specific goals, and such files are not used by other files. The evidence is consistent with corporate contributions being aimed at the intensive margin: More intense drives of effort. While corporate created files attract non-corporate individuals' effort as well, the type of file is important. For files that are hardware specific, we see greater non-corporate contributions, but for general files, we see no difference between corporate and non-corporate contributions. Similarly, for corporate authored generic files, we observe less follow-on innovation.

We complement our analysis with causal evidence from the emergence of the Open Handset Alliance (OHA) which increased corporate commercialization incentives. Our difference-in-difference estimates comparing corporate and non-corporate authored files before and after OHA show that increased corporate commercialization activities can crowd out individual contributions. This effect is pronounced for generic files.

We contribute to the literature on innovation in crowdsourced digital goods. In particular, we contribute to a growing literature documenting how crowdsourced digital goods evolve (Kane and Ransbotham 2016, Nagaraj 2017). We uncover an additional source of limitations to crowdsourced innovations that originates in different incentives facing individuals and corporations.

This suggests the potential necessity of government participation in maintaining the exploratory nature of crowdsourced digital goods. We also contribute to the literature on the benefits from contributing to, and using open source software (Lerner and Tirole 2002, Nagle 2018a, b). Finally, we contribute to the literature on innovations and private/public streams of knowledge (Aghion et al. 2008, Murray and Stern 2007, Stern 2004).

References

Aghion P, Dewatripont M, Kolev J, Murray F, Stern S (2010) The public and private sectors in the process of innovation: Theory and evidence from the mouse genetics revolution. *American Economic Review* 100(2):153–58.

Aghion P, Dewatripont M, Stein JC (2008) Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics* 39(3):617–635.

Ahuja G, Lampert CM, Novelli E (2013) The second face of appropriability: Generative appropriability and its determinants. *Academy of Management Review* 38(2):248–269.

Baker G (2000) The use of performance measures in incentive contracting. *American Economic Review* 90(2):415–420.

Belenzon S, Schankerman M (2015) Motivation and sorting of human capital in open innovation. *Strategic Management Journal* 36(6):795–820.

Boudreau K (2010) Open platform strategies and innovation: Granting access vs. devolving control. *Management science* 56(10):1849–1872.

Casadesus-Masanell R, Llanes G (2011) Mixed source. *Management Science* 57(7):1212–1230.

Corbet J, Kroah-Hartman G (2017) *2017 Linux Kernel Development Report* (The Linux Foundation).

Gambardella A, von Hippel EA (2018) Open source hardware as a profit-maximizing strategy of downstream firms.

Greenstein S, Nagle F (2014) Digital dark matter and the economic contribution of Apache. *Research Policy* 43(4):623–631.

Henkel J (2006) Selective revealing in open innovation processes: The case of embedded Linux. *Research policy* 35(7):953–969.

Henkel J, Schöberl S, Alexy O (2014) The emergence of openness: How and why firms adopt selective revealing in open innovation. *Research Policy* 43(5):879–890.

Von Hippel E, Von Krogh G (2003) Open source software and the "private-collective" innovation model: Issues for organization science. *Organization science* 14(2):209–223.

Huang KG, Murray FE (2009) Does patent strategy shape the long-run supply of public knowledge? Evidence from human genetics. *Academy of management Journal* 52(6):1193–1221.

Kane GC, Ransbotham S (2016) Content as community regulator: The recursive relationship between consumption and contribution in open collaboration communities. *Organization Science* 27(5):1258–1274.

Kumar V, Gordon BR, Srinivasan K (2011) Competitive strategy for open source software. *Marketing Science* 30(6):1066–1078.

Lakhani KR, Von Hippel E (2003) How open source software works:"free" user-to-user assistance. *Research policy* 32(6):923–943.

Lerner J, Schankerman M (2010) The comingled code: Open source and economic development. *MIT Press Books* 1.

Lerner J, Tirole J (2002) Some simple economics of open source. *The journal of industrial economics* 50(2):197–234.

Leroux S (2017) Open Source is Taking Over Europe! - It's FOSS. *https://itsfoss.com/*.

Llanes G, de Elejalde R (2013) Industry equilibrium with open-source and proprietary firms. *International Journal of Industrial Organization* 31(1):36–49.

MacCormack A, Rusnak J, Baldwin CY (2006) Exploring the structure of complex software designs: An empirical study of open source and proprietary code. *Management Science* 52(7):1015–1030.

Murphy GC, Notkin D, Griswold WG, Lan ES (1998) An empirical study of static call graph extractors. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 7(2):158–191.

Murray F (2010) The oncomouse that roared: Hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. *American Journal of sociology* 116(2):341–388.

Murray F, Aghion P, Dewatripont M, Kolev J, Stern S (2016) Of mice and academics: Examining the effect of openness on innovation. *American Economic Journal: Economic Policy* 8(1):212–52.

Murray F, Stern S (2007) Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization* 63(4):648–687.

Nagaraj A (2017) Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap. *Available at SSRN 3044581*.

Nagle F (2018a) Learning by Contributing: Gaining Competitive Advantage Through Contribution to Crowdsourced Public Goods. *Organization Science*.

Nagle F (2018b) Open source software and firm productivity. *Management Science*.

Parker GG, Van Alstyne MW (2005) Two-sided network effects: A theory of information product design. *Management science* 51(10):1494–1504.

Roach M, Sauermann H (2010) A taste for science? PhD scientists' academic orientation and self-selection into research careers in industry. *Research policy* 39(3):422–434.

Simcoe T (2012) Standard setting committees: Consensus governance for shared technology platforms. *American Economic Review* 102(1):305–36.

Stern S (2004) Do scientists pay to be scientists? *Management science* 50(6):835–853.

Von Hippel E (1986) Lead users: a source of novel product concepts. *Management science* 32(7):791–805.

Wasko MM, Faraj S (2005) Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly* 29(1):35–57.

West J (2003) How open is open enough?: Melding proprietary and open source platform strategies. *Research policy* 32(7):1259–1285.

West J, Gallagher S (2006) Challenges of open innovation: the paradox of firm investment in open-source software. *R&d Management* 36(3):319–331.

Zhang XM, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101(4):1601–15.

Appendix A: Appendix to Chapter 1

Section 1. Generalizability

1.1 Qualitative Case Studies Related to Phenomenon of Interest

Table A1. Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
1	Migration of Huguenots to Brandenburg-Prussia	Huguenots (French Protestants) migrated to Brandenburg-Prussia (Northern German states) from 1685 till 1789. The migration was motivated by religious persecution in France and religious protection policies in Brandenburg-Prussia.	Many Huguenots distinguished themselves as skilled artisans and craftsmen, especially cloth-workers, and bore specialized and "secret" knowledge related to these crafts. One account "provides a list of 46 professions introduced by Huguenots to Brandenburg, all of which were previously unknown to the country, most of them in the textile industries. One Huguenot carried with him the secret of dyeing fabrics in a special way, another brought the art of printing on cotton (Hornung 2014, 91, 93).	In one example, Huguenots transmitted knowledge about silk production into Brandenburg-Prussia (Hornung 2014, 94). This stimulated a surge in the planting and cultivation of one preexisting crop in Brandenburg-Prussia: mulberry plants. There was a general impression that silk worms cannot thrive in the cold temperatures of the northern German states, such as Prussia. To help overcome this, concurrent with the influx of Huguenot immigrants, rulers in the northern German states began to promote the planting of mulberry tree plantations in their territories, the crop of which was used to feed silkworms. The historical record notes that "it is schoolmasters who chiefly occupy themselves" with the planting of silkworms, as a means of "adding to [their] income." The rule of Brandenburg-Prussia, who had encouraged Huguenot immigration, is noted to have "ordered the cultivation of mulberry trees in schoolyards to feed the silkworms." The resultant quality of silk in the northern German countries has been noted as "remarkably white, and finer than that in the southern countries (Scientific American 1853).

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
2	Migration of Soviet mathematicians to the United States	Mathematicians from the Soviet Union migrated to the United States between 1990 and 2000, after the fall of the Soviet Union. In 1992, the Soviet Scientists Immigration Act was passed into law in the United States, which allocated visas to be given to Soviet Scientists to immigrate to the U.S. As Ganguli (2015) states, estimates from the 2000 Census suggest that close to 10,000 Russian scientists and engineers across many science and technology fields immigrated to the United States in the 1990s.	During the Soviet era, Soviet mathematicians worked in mathematical knowledge sub-fields that differed from those Americans worked in (Borjas and Doran 2012, 6-9). Borjas and Doran (2012) show that Russian mathematicians were ahead of the west in fields like partial differential equations and symplectic topology. Such knowledge was also arguably ex ante "locked" within the soviet context, prior to the migration of Soviet scientists to the United States. Borjas and Doran (2012) outline two reasons for why such knowledge was locked within the Soviet context. First of all, as Abramitzky and Sin (2011) report, the translation rate of hard-science Eastern Bloc books into English was extremely low. Secondly, Borjas and Doran (2012) cite examples from Tybulewicz (1970) to state that even if translations of Soviet academic work were available, the knowledge was often not transferred to American academics, because they did not read the translations of Soviet scientific work.	An influx of Soviet mathematicians (roughly 300 during the 1990s) immigrated to the U.S. after the Soviet regime waned and collapsed in the early 1990s (Borjas 2014, 183). While they competed with American mathematicians for jobs and publication space in journals, they also collaborated and helped American mathematicians solve formerly intractable problems (Borjas and Doran 2012, 26, 11). One study cited news coverage from the time (from the New York Times) which described how "[American mathematician] Dr. Diaconis said he recently asked [Soviet mathematician] Dr. Reshetikhin for help with a problem that had stumped him for 20 years. 'I had asked everyone in America who had any chance of knowing' how to solve a problem . . . No one could help. But . . . Soviet scientists had done a lot of work on such problems. 'It was a whole new world I had access to,' Dr. Diaconis said. 'Together, we'll be able to solve the problem.

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
3	Migration of skilled artisans from Florence to Venice (driven by patent laws) and the influence of such migration on Silk Dyeing in Venice	Belfanti (2004) documents an interesting case study where migration of skilled artisans in Italy between the sixteenth and eighteenth century was driven by patent laws. As Belfanti (2004) documents, the city council and princes of Italian Republics such as the Venetian Republic used the attractiveness of assigning monopoly rights to intellectual property through patents, to attract skilled migrant workers from other regions, such as Florence. An interesting example related to this phenomenon dates back to the eighteenth century, when skilled silk dyers migrated from Florence to Venice.	While artisans from Venice possessed deep knowledge in making Brocade, they lacked knowledge in the area of silk dyeing. To transfer knowledge related to silk dyeing from Florence to Venice, as Belfanti (2004) documents, at the beginning of the eighteenth century, emissaries of Venice traveled to Florence to recruit silk dyers—an effort that succeeded in convincing dyers such as Cosmo Scatini, to migrate to Venice. Scatini was a Florentine dyer who knew the secret of dyeing silk black. Between 1727 and 1732 the Venetians also twice tried to bring experts in the making of silk veils from Bologna. To attract migrant artisans who possessed specialized knowledge to migrate to Venice, the Venetian Republic awarded them with patents. Cosmo Scatini, the migrant from Florence, was awarded with a patent for skill dyeing. Another migrant artisan craftsman who introduced the weaving of silk stockings on a frame was rewarded with a patent and a ten-year monopoly.	While the Venetian Republic used patents to attract skilled migrants from Florence and elsewhere, there was also a strong guild of local artisans in Venice, and the incentives of local artisans were geared towards working on the patented technology, at the end of the patent term. While the patents awarded to skilled migrant artisans granted them monopoly rights for a period of time, at the end of the fixed monopoly term, the migrant artisans were required to share their knowledge with the local guild of artisans. As Belfanti (2004) says, the Florentine artisan Cosmo Scatini, who obtained a patent for silk dyeing, applied to be enrolled in the dyers' guild once his privilege ran out, promising to teach the Venetian craftsmen the process. This transfer of knowledge often led to knowledge recombination. An example of such knowledge recombination relates to dyeing Black silks in Venice. Black was a symbol of superiority in Venice in the late 17th century (Bervegliari 1983, 176). When Cosmo Scatini transferred his secret knowledge of black silk dyeing to Venetians at the end of his patent term, the dyeing techniques changed in Venice and they were able to produce black silks (Bervegliari 1983, 176).

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
4	Migration of Italians to United States	As Choate (2008) documents, between 1880 and 1924, more than four million Italians immigrated to the United States, half of them between 1900 and 1910 alone—the majority fleeing grinding rural poverty in Southern Italy and Sicily. In the period 1880 to 1915, total Italian emigration has been estimated at 13 million, making it the largest emigration from any country in recorded history.	Natives of Campania and Sicily nurtured a cuisine of the tomato, onion, oil, cheese, and garlic. As Levenstein (1985) documents, the preparation and consumption of food was central to Italian family life. Italian women in particular "retained" knowledge about "community-borne recipes and instructions in cooking." (Levenstein 1985, 76, 80). Knowledge about *how* to cook with key ingredients such as these was tacit knowledge possessed by families, and was transferred to the United States by the migrants.	Italians in America resisted assimilation in general, and in particular resisted "Americanizing" their own cooking habits. A type of distinctive Italian cuisine in America thus took hold and key dishes became popular in American households after World War I. To bring basic Italian dishes and cooking methods to American households, such as "spaghetti in tomato sauce," brands such as Campbell's and Heinz marketed shelf-stable versions (Levenstein 1985, 79-81, 86). In doing so, they applied their own knowledge of processing and packaging food to traditional Italian recipes. As Levenstein (1985) documents, this recombination is in stark contrast to the negative perception towards ingredients such as tomatoes, traditionally harbored by American society. In the late 1880s, prior to the migration from Campania, the scientific community in the United States held a belief that tomatoes were carcinogenic, and were generally harmful due to the presence of "oxalic acid". The Italian migration, the transfer of knowledge related to Italian ingredients to the United States and the subsequent recombination of such knowledge went a long way to allay prior scientific "beliefs".

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
5	Ethnic migrants and their influence on the garment industry in New York	The New York garment industry has been shaped by waves of ethnic migrants from Germany, Ireland, Russia (Russian Jews), Italy, China, and Puerto Rico.	Till the late 1800s, different ethnicities migrating to New York transferred knowledge of their distinctive sartorial designs to their host region. As Bagger (1871) states, in the mid-1800s in New York, an individual's nationality could be determined by how he or she dressed. To quote the author, "It is curious to see such a heterogeneous crowd land. The Swedes are usually distinguished by their tanned-leather breeches and waistcoats, and their peculiar before mentioned exhalations; you can not miss the Irishman with his napless hat, worn coat, and corduroy trousers; the Englishman you know by his Scotch cap, clay pipe, and paper collar".	Among other scholars, Rantisi (2002) documents subsequent recombination of knowledge transferred by ethnic migrants, in the context of the New York garment district. The garment district in New York was a giant melting pot for ethnicities to work together and for knowledge recombination. As Rantisi (2002) states, while the origins of the New York garment industry can be traced to German Jews, migration of Italians in the 1880s and Russian Jews later led to knowledge recombination. In an online article, Dzvinka Stefanyshyn provides examples of ethnic migrant influence on knowledge transfer and knowledge recombination. To quote the author, "when the Italian and Jewish immigrants dominated clothing manufacturing businesses, certain aspects of style were altered. Up until the 1900s, women wore tight-fitting clothing, which included uncomfortable corsets and long, extravagant dresses. Turning away from that fashion, the workers focused on creating clothing that was much simpler and comfortable. This resulted in the introduction of much looser, natural – either woolen or linen – clothing".

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
6	Migrants from England to Italy (the silk machines example)	Temporary migration from England to Italy by migrants such as John Lombe in 1607, presumably for transferring knowledge related to "silk throwing".	In 1607 Vittorio Zonca published in Padua *Nuovo Teatro di Machine et Edificii* describing a machine to throw silk by water power. While this book was available in the UK, providing the know-why, it was only after a temporary migrant, John Lombe traveled to Italy and spent 2 years learning the know-how, the knowledge of silk throwing transferred to the UK (Cipolla 1972). To quote the author, "In 1607 Vittorio Zonca published in Padua his *Nuovo Teatro di Machine et Edificii* which included, among numerous engravings of various machines, the description of an intricate machine for throwing silk by water power in a large factory….notwithstanding the description by Zonca, the details of the mill were still considered state secret and Piedmontese laws regarded 'the disclosing or attempting to discover' anything relating to the making of the engines a crime punishable by death. The Piedmontese were no fools. G. N. Clark has pointed out that a copy of the first edition of Zonca's book had been on the open-access helves of the Bodleian Library from at least as early as 1620. Yet the English succeeded in building a mill for the throwing of silk only after John Lombe, during two years of industrial espionage in Italy, 'found means to see this engine so often that he made himself master of the whole invention and of all the different parts and motions" (Cipolla, 1972, page 47)	When Lombe returned from Italy to Britain, he brought with him first-hand know how of using of silk machines, and workers who could help set up the factories (Calladine 1993) to create British silk machines.

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
7	Italian migrants in Australia	As Castles (1992) documents, mass migration from Italy to Australia happened between 1950-1970; between 1951-1961, 18,000 new migrants arrived every year.	Genovesi (2000) documents the transfer of knowledge related to food and fashion by Italian migrants in Australia. To quote the author, "One needs to understand that Italian migrants needed particular foods for the practicing of social events, cultural and catholic religious rituals. For example, sugar almonds were needed for the rituals of the sacraments, dried fish for Good Friday, "la castagnata" (chestnuts) Christmas sweets, essence for cakes, alcohol such as wine and grappa. Amongst the Italians, there was little tolerance for mutton, fish and chips, or the meat pie". The author also documents transfer of knowledge across borders. To quote, "Many Italian immigrants felt the need to expand on the availability of vegetables and fruits and took matters into their own hands…..They had brought with them skills to make pasta, breads, sauces, preserves, cheeses, wines and pork sausages. In addition, as the immigrants arrived, so did particular foods, seeds, kernels and cuttings, mostly as contraband. In my conversations within the Italian community, it has been stated that these items were hidden in suitcases, coat linings, pockets and underwear." (Genovesi, 2000; page 8).	Over time, Italian food (and fashion) has experienced recombination in Australia. An example relates to the dish, Chicken Parmigiana. One account describes the dish as "An Italian name, but a bona fide Australian pub classic, the parmigiana started as an eggplant dish in Italy and has since evolved into a chicken schnitzel topped with an Italian-inspired tomato sauce and melted cheese." (source: http://www.cnn.com/travel/article/australian-food/index.html). Another source describes how the original recombination was within the Italian migrant communities in the United States and then the recombined dish became popular within Italian migrants in Australia. To quote, "Chicken Parmigiana has its origins in the United States, where it was popularized among Italian-American communities. Italian immigrants created the meal, which quickly became conceived as an authentically Italian dish. Of course, it does take its inspiration from Italy. Eggplant Parmigiana, or Mellenzana alla Parmigiana, is the original Italian recipe. Eggplants are lightly breaded, fried, topped with fresh tomato sauce and Parmesan cheese, and then baked. The switch to chicken in the United States might have been due to several reasons – Italian restaurant owners saw the American preference for meat over eggplant, Italian immigrant workers were able to afford meat now that they had higher paying jobs, or eggplants just weren't as common a produce in the United States." (source: https://www.montebene.com/blogs/blog-posts/58998467-the-story-behind-the-staple-chicken-parmigiana)

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
8	Migration of Indians to South Africa and East Africa and the impact of such migration on the transfer or accounting practices across borders	Indian merchants from Gujarat migrated to Natal, South Africa and to Kenya between 1875 and 1910 in search of better economic opportunities. This example documents the transfer of knowledge related to accounting practices from their home region to their host region, and subsequent recombination of such knowledge at their host region.	The migrants transferred knowledge related to an accounting practice called "Hundi" from Gujarat, India to Africa. A Hundi is most often translated as a "bill of exchange" or "promissory note." It functions as both a credit system and a remittance system, and historically has been linked to the Indian merchant community conducting trade in the Indian Ocean region. The bookkeeping system associated with the hundi exchange is known as *hundi khata*: a system which emphasizes "double-entry" accounting, and thus differs from the traditional English "bills book" system (Nigam 1986, 148, 156). The knowledge of this accounting practice was locked within the Gujarat region of India, because it was documented in vernacular text. The British Museum has referenced an autobiographical poem published in the 17th century by the Gujarati (Jain) poet and businessman Banarasidas as containing one of the earliest known references to hundi and Martine (2009) documents that the accounting practice was used exclusively within Gujarat because the creditors used the local vernacular language to write the promissory note.	Post the migration of Gujarati traders to Africa, the hundi system flourished in East Africa. European banks in the region also began to encourage the use of the "overdraft" as a medium for credit lending to "reliable" customers. The combination of the hundi and overdraft systems contributed to the development of a loosely-defined "chit" system in East Africa, as distinct from lending and accounting conventions popular in England at the time (Gregory 1993, 103). Prior research has documented that this recombined system allowed individuals to write checks or withdraw currency well beyond the amounts of their deposits on the understanding that, when able, they would make up the difference and pay a small interest on the overdrafts. As Gregory (1993) states, East Africa as early as 1907 was becoming known as 'the land of the chit.'" Notably, the English word "chit"—defined as "a signed voucher of a small debt"—itself originated from Hindi and Urdu, languages in Northern India (See Merriam-Webster, "Chit," accessed September 11, 2017, https://www.merriam-webster.com/dictionary/chit.)

Table A1. (Continued) Qualitative Case studies

#	Qualitative Example	Details of Migration	Knowledge Being Transferred Across Borders	Subsequent Recombination of Knowledge
9	Migration of Chinese to the United States and the creation of "American Chinese" food	Chinese migrants moved to the US because of the California Gold Rush in the 1850s. The early Chinese migrants were Cantonese from Guangdong area who entered primarily through port of San Francisco (source: The Search for General Tso). The Chinese Exclusion Act was signed in 1882, blocking Chinese immigration and naturalization, which led to attacks and threats on those already living in the US (Coe 2009). Chinese settlers began to move out of California and spread across the US to escape persecution and search for employment opportunities.	The documentary film The Search of General Tso documents the transfer of knowledge related to food from China to the U.S. When Chinese migrants first arrived to the U.S., Americans were both fascinated and repulsed by their food. After the signing of the 1882 Chinese Exclusion Act, many Chinese migrants were forced out of labor and had to be self-employed, occupying two main industries: laundry and food (source: The Search for General Tso). Chinese migrants began opening restaurants across the US. General Tso's chicken, arguably the most popular Chinese food dish in America, was brought to the U.S. from Taiwan, where former Hunan chef Peng Chang-kuei had fled after the Communist Revolution of 1949 (Bateman 2016). It was there that he created General Tso's chicken, inspired by the spicy and sour Hunan palate. Peng moved to New York City in 1973 and some believe he brought General Tso's chicken with him (Bateman 2016). It is also argued that chefs from New York visited Taiwan and brought Peng's dish with them prior to his move to the U.S.	Following the 1882 Chinese Exclusion Act, "restaurant work was one of the few jobs that the Chinese could find, and Chinese restaurant owners discovered that if they adapted simple dishes to American taste, that they could make money" (source: The Search for General Tso). According to the film, chefs began to change menus based on demand, and ultimately adapted their food to appeal to white audiences By 1900, the popular dish "chop suey" had been created, combining "Americanized meats and 'exotic' flavorless vegetables," and a mixed menu of American food and Chinese American food became national phenomenon (source: The Search for General Tso). In 1940, David Leong moved to Springfield, Missouri and created cashew chicken to appeal to the local population, frying the chicken, knowing that American's in the south loved fried chicken. Fortune cookies were also invented in the US in the 1940s, unique to American Chinese cuisine. Different regions adapted Chinese food to fit the local demand. In Louisiana, they serve Chinese gumbo with alligator meat. General Tso's Chicken became an American phenomenon in the 1970s when Michael Tong and chef T. T. Wang opened the Shun Lee Palace in NY. Wang takes credit for inventing the dish, but he was actually inspired by Peng Chang-kuei's invention in Taiwan, and added more sugar because he thought the American palate was sweeter than the Chinese (source: The Search for General Tso).

1.2 Links to Traditional Scientific Research

We also document the extent of herbal ingredients in the science literature. Using our list of herbs, we search for scientific articles listed on PubMed containing any of our herbs as dietary supplements. Towards this, we utilized the PubMed Dietary Supplements Subset. The PubMed Dietary Supplements Subset allows researchers to retrieve dietary supplement related citations on topics such as traditional Chinese medicine and herbal medicine, among other topics. Queries of herbal ingredients will return dietary supplement related articles on PubMed that contain the queried ingredients. Searches using our herbs resulted in 658,488 articles on PubMed, published in 11,974 unique scientific journals.

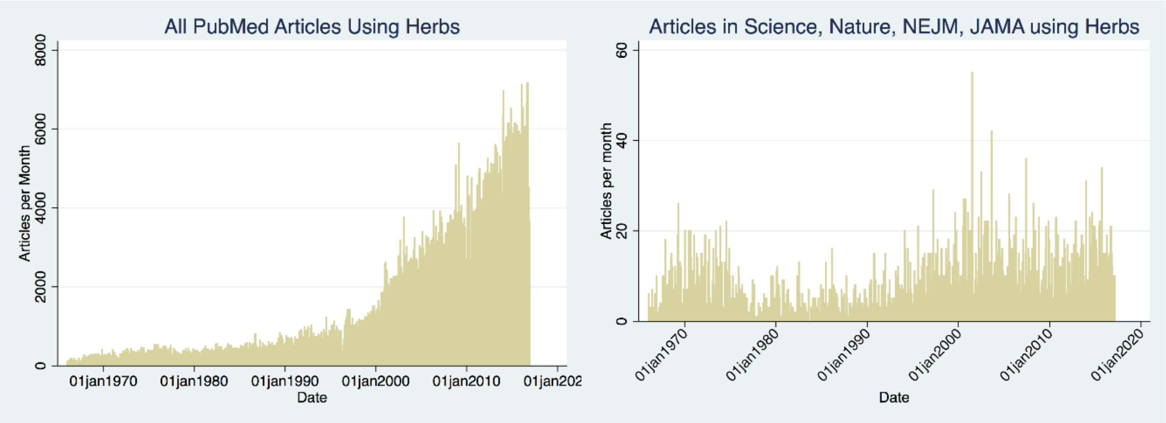


Figure A1. *Articles in scientific research using herbal supplements* Frequency plots of scientific articles on PubMed over time. We see a general increase in articles using herbal supplements over time. Restricting the subset to the most impactful journals also show a general increase in herbal research.

Table A2. NBO Company Shares of Herbal/Traditional Products: % Value 2012-2016

% retail value rsp	2012	2013	2014	2015	2016
Mondelez International Inc	7.8	7.8	7.7	7.4	6.9
Procter & Gamble Co, The	3.4	3.0	3.1	3.0	2.9
Ricola Inc	2.7	2.9	2.9	3.0	2.9
GSK Consumer Healthcare	-	-	-	1.3	1.5
Prestige Brands Inc	1.3	1.3	1.3	1.3	1.3
McNeil Consumer & Specialty Pharmaceuticals	1.3	1.2	1.1	1.0	1.0
NBTY Inc	1.0	1.0	0.9	0.9	0.8
NFI Consumer Products	0.2	0.3	0.5	0.7	0.8
Herbalife International Inc	0.8	0.9	0.8	0.8	0.8
General Nutrition Centers Inc	0.8	0.8	0.8	0.8	0.7
Forever Living Products LLC	0.8	0.8	0.8	0.7	0.7
Korea Ginseng Corp	0.4	0.4	0.5	0.6	0.7
Haw Par Healthcare Ltd	0.5	0.5	0.5	0.6	0.6
CNS Inc	0.9	0.7	0.7	0.6	0.6
Amway Corp	0.6	0.6	0.6	0.6	0.6
Nature's Way Products Inc	0.4	0.5	0.5	0.5	0.5
Performance Health Inc	0.4	0.4	0.5	0.5	0.5
Chattem Inc	0.6	0.5	0.5	0.4	0.4
Perfecta Products Inc	0.2	0.3	0.3	0.4	0.4
Nutraceutical International Corp	0.3	0.3	0.3	0.4	0.4
Wakunaga Pharmaceutical Co Ltd	0.4	0.4	0.4	0.4	0.3
Lily of the Desert Organic Aloeceuticals	0.3	0.3	0.3	0.3	0.3
Nature's Sunshine Products Inc	0.3	0.3	0.3	0.3	0.3
Troy Healthcare LLC	0.3	0.3	0.2	0.2	0.2
Pfizer Consumer Healthcare Inc	0.3	0.2	0.2	0.2	0.2
Concepts in Health	0.5	0.4	0.3	0.3	0.2
Windmill Health Products	0.2	0.2	0.2	0.2	0.2
DSE Healthcare Solutions LLC	-	0.2	0.2	0.2	0.2
Alan James Group LLC	0.2	0.2	0.2	0.2	0.1
Smith Bros Co, The	0.1	0.1	0.1	0.1	0.1
Novartis Corp	0.7	0.5	1.0	-	-
WF Young Inc	0.2	-	-	-	-
Other Private Label	0.6	0.6	0.6	0.5	0.5
Others	71.5	72.0	71.5	71.8	72.4
Total	100.0	100.0	100.0	100.0	100.0

Source: Euromonitor International from official statistics, trade associations, trade press, company research, store checks, trade interviews, trade sources

Section 2. Validity of the visa shock

2.1 Labor Condition Applications

One concern is that the H-1B visas were targeted towards large IT companies, and firms writing herbal patents would not have been affected by the visa shock. While we are not able to observe the number of H-1B visa grants at the firm level, we can observe the number of Labor Condition Applications (LCAs), a prerequisite of H-1B visas, applied for by each firm. In this section, we argue that herbal patent assignees benefit more from H-1B visas than the average firm does. Back of the envelope calculations show that the visa shock allowed them to hire 424 Chinese/Indians instead of the 141 under stricter immigration laws, and led to 68 additional patents. These numbers are consistent with the patenting and hiring rates of Amgen, suggesting our calculations are plausible.

The Foreign Labor Certification Data Center provides historical data on LCAs issued since 2000. We take all 633 herbal patent assignees located in the US that filed for herbal patents, and match them to a list of entities that filed LCAs between 2000 and 2016. This includes all U.S. based assignees active during the visa shock period who have filed a herbal patent; in some cases the herbal patents filed may not have been granted. In fact, the sample size drops to 401 assignees if we consider only assignees who have been granted a herbal patent. Using a fuzzy string matching algorithm, we calculate all pairwise string similarity scores between patent assignees and LCA filing entities, and manually inspect matches to create a cutoff score above which we will consider entities to be a match. (In the following discussion, we use a score of 93 as the cutoff, but the results are similar to using cutoffs of 94 or 95).

We next show herbal patent assignees file more LCAs than other firms on average, and thus are likely to have more H1B hires than the average firm. We use the matched sample from above and plot the quantile-quantile plot of total LCAs filed by our assignees and all other firms (Figure A2). The quantile-quantile plot does not follow the 45 degree line, implying the two distributions are different. Furthermore, we see the number of LCAs filed by herbal patent assignees is left skewed, suggesting that herbal assignees are more likely to hire more people through the H1B visas than other firms filing LCAs. T-test results show that herbal assignees file for 146.7 more LCAs (*t*-statistic 4.81) than other firms, further showing herbal patent assignees are a major beneficiary of the H-1B visas.

Going forth, we make three assumptions that will allow us to measure the relationship between hiring ethnic migrants and patenting. First, we assume LCA filings for treated assignees are directly correlated with their H1B grants, and because 0.1666% of LCA filings are by treated assignees, they will collect 0.1666% of the H1B visas. Second, we assume that 40% of H1B visa grants during 1999-2003 go to inventors with Chinese/Indian nationalities. Our third assumption is that we can estimate the number of patents granted to treated firms during the shock period by predicting counterfactuals using our difference in difference specification.

Our calculations suggest the tripling of the visa cap allowed firms to hire 283 more Chinese/Indian inventors, in addition to the 141 they would have hired under stricter immigration laws. We arrive at this conclusion as follows. The total number of LCAs filed between 2001 and 2016 is 25,128,680, and the number of LCAs filed by capped herbal patent assignees is 41,874. This shows that capped herbal patent assignees file 0.1666% of all LCAs. If capped herbal patent assignees secured H-1B visas in the same proportion as the filed LCAs, this suggests that of the 636,994 H-1B visas issued during the visa shock period, 1999-2003, our firms would have used around 1,061. If the fraction of Chinese/Indian inventors is around 40%[71] (an assumption borrowed from a report filed by the Center of Immigration Studies or CIS), herbal patent assignees would have hired 424 new Chinese/Indian inventors[72]. Roughly two thirds, or 283 inventors, would not have been available for hire under the stricter visa regulations.

We next show around 17.45 percent of these inventors are granted patents, and this leads to roughly 68 new patents (or 0.1604 patent per Chinese/Indian H1B hire). During the visa shock period, we observe 74 new Chinese/Indian inventor names in our patent data, suggesting about 17.45 percent of new hires file and are granted patents (we get this by dividing the 74 new Chinese/Indian names that we observe in the patent records, by the total of 424 new Chinese/Indian inventors that we estimate were hired under H-1B at the capped firms). We further calculate a rough estimate of the number of additional herbal patents filed by predicting patent counts with our fitted regression models. First, we obtain predicted patent counts by plugging in our data into the model with time and assignee fixed effects (column 4 of Table 2) and exponentiating the results. After obtaining the predicted patent counts, we predict a counterfactual log patent count by setting the coefficient on the interaction to zero, again plugging in our data and exponentiating, and take the difference of the two predictions. Summing up the differences, we see that the visa shock generated an additional 68.06 patents, or 0.1605 patents per Chinese/Indian H1B hire. Finally, we show that the estimates above align with a well-known pharmaceutical firm's H1B hiring and patenting activities. We are able to observe H1B hires and patenting for Amgen in 2015[73]. In

[71] CIS documents show that in 2005, the fraction of Chinese/Indian H-1B beneficiaries was less than 45%: https://www.cis.org/Report/Wages-H1B-Computer-Programmers

[72] We only observe new inventors for granted patents, and around 63% of assignees that file herbal patents are granted one.

[73]https://www.uscis.gov/sites/default/files/USCIS/Resources/Reports%20and%20Studies/Immigration%20Forms%20Data/BAHA/h-1b-2015-employers.pdf

2015, they filed for 420 LCAs, 80 of which led to H1B visas. In addition, because 80% (note that the percentage of H1Bs granted to Chinese/Indians increases in 2015 compared to the 2005 fraction according to the USCIS reports) of H1Bs were granted to Chinese/Indians in 2015, around 64 of these hires would be Chinese/Indian nationals. We calculate the number of herbal patents by Chinese/Indians by counting the number of herbal patents by Amgen in 2015 in our data. We observe Amgen filed 71 herbal patents that year, 9 of which were granted. This suggests that 1 additional Chinese/Indian inventor amounts to 0.1406 herbal patent grants, an estimate similar to the 0.1605 obtained above.

Thus, our back of the envelope calculations suggests that an increase in the visa cap led to 424 new hires of Chinese/Indian ethnicity by herbal patent assignees, 74 of which (around 17.45 percent) were granted a patent during the same period. Hiring these inventors led to an increase of 68 herbal patents for our assignees. Hiring patterns and patenting patterns of a large firm are consistent with our calculations, adding to the credibility of our methods.

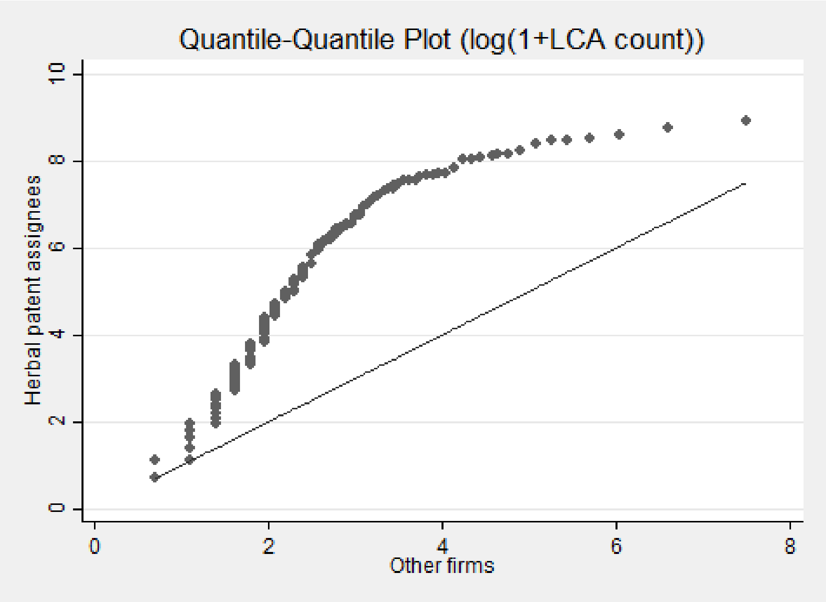


Figure A2. Quantile-Quantile plot of herbal patent assignees and all other organizations that filed for an LCA

2.2 Visa shock increased Chinese/Indian inventor hiring

We further test whether the visa shock increased the likelihood of hiring new Chinese/Indian inventors. Towards this, we estimate the following regression equation

$$1(HiredEthnic)_{jt} = \alpha + \beta_1 Capped_j + \beta_2 Shock_t + \gamma Capped_j \times Shock_t + \phi_j + \lambda_t + \varepsilon_{jt}$$

Our main dependent variable is an indicator variable for whether firm j hired any Chinese/Indian inventors at time t. We present the results in Table A3.

Table A3. Effect of visa shock on hiring Chinese/Indian inventors

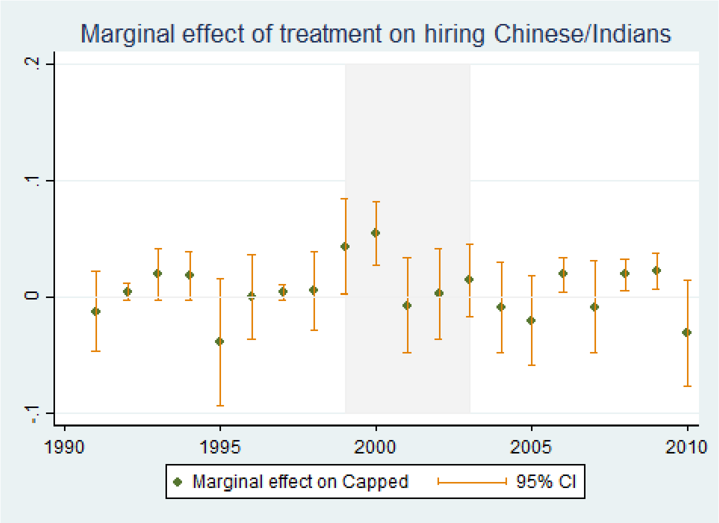
	(1)	(2)	(3)	(4)
	Dep Var: 1(Hired Chinese/Indian inventor)			
Capped x Shock	0.02	0.02	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)
Capped	-0.00	0.00	-0.00	
	(0.00)	(0.00)	(0.00)	
Shock	0.01	0.01		
	(0.01)	(0.01)		
Constant	0.01	-0.00	-0.02	-0.06
	(0.00)	(0.00)	(0.00)	(0.00)
Controls	N	Y	Y	Y
Time FE	N	N	Y	Y
Firm FE	N	N	N	Y
Observations	8998	8998	8998	8998
Adjusted R^2	0.005	0.039	0.041	0.021

Note: Standard errors in parentheses, clustered at the assignee level. Observations at the assignee-year level, for all years an assignee (firm or University) was in operation. We use the tsfill command in Stata to fill in missing assignee-year pairs. Dependent variable is an indicator for whether the firm hires a Chinese or Indian inventor in a year. Controls include the fraction of Chinese/Indian inventors, firm age, inventor count, and total number of Chinese and Indian inventors. *p*-values for the interaction term (row 1) across columns are 0.030, 0.045, 0.027, and 0.028 respectively.

In the baseline specification, we see that firms subject to the visa cap increased hiring of Chinese/Indian inventors by 1.81 percent, compared to exempt firms. The p-value for this coefficient is 0.03, again suggesting that these effect sizes are highly unlikely under the null hypothesis of no effect. Compared to the baseline likelihood 1.06 percent of hiring a Chinese/Indian inventor, we see the visa shock increased the likelihood of hiring a Chinese/Indian inventor by 171 percent. The effect size is similar after including controls (Column 2), time fixed effects (Column 3), and firm fixed effects (Column 4). The p-value for our final result is 0.028, and the coefficient suggests that treatment increased the likelihood of hiring a Chinese/Indian inventor by 187 percent compared to the baseline. Interestingly, we do not see large differences in the time invariant likelihood of hiring Chinese/Indian inventors across capped and exempt firms, nor in the time effect (second and third rows respectively).

We repeat the lead-lag analysis as above, and plot the resulting coefficients in Figure A3. Again, the 95% confidence intervals always include zero before the shock, suggesting that there are no pre-trends. Here we see the effect of the visa shock on hiring over time as well. While in 1999 and 2000 the confidence intervals do not include zero, from 2001 onwards we fail to reject a null effect. This echoes the fact that the visa quota was met in 1999 and 2000, but not in 2001-2003[74].

[74] https://redbus2us.com/h1b-visa-cap-reach-dates-history-graphs-uscis-data/



A3. Leads and lags for DD specification for hiring Chinese/Indians

Section 3. Ethnic inventors bring in new knowledge

In this section, we test whether new inventors, and in particular new ethnic inventors, are more likely to introduce knowledge that was previously geographically locked. One implication of knowledge that is previously locked is that it would be less familiar in the US context, in our case measured by whether a patent's herbs are new, and how familiar the herb is (as measured by Google Ngrams). Then, a testable hypothesis is whether new inventors, in particular new Chinese/Indian inventors, are more likely to introduce herbs that are less familiar in the US context. We test this hypothesis by estimating the following regression equation using ordinary least squares:

$$\boldsymbol{New\ Herb_{ijt} = \beta_0 + \beta_1 New\ Inventor_{ijt} + \beta_2 New\ Ethnic\ Inventor_{ijt} + \phi_j + \lambda_t + \varepsilon_{ijt}}$$

We use two measurements for our $New\ Herb_{ijt}$ variable. In the first case, it is an indicator for whether patent i introduces a new herb at time t. In the second case, it measures the log frequency (as measured by Google N-grams) of the newly introduced herb used in patent i, applied for by firm j at time t. $New\ Inventor_{ijt}$ is an indicator for whether patent i includes a first-time inventor, and $New\ Ethnic\ Inventor_{ijt}$ is an indicator for whether patent i includes a first-time inventor who also has a Chinese or Indian name. We include firm fixed effects ϕ_j, and time fixed effects λ_t to control for firm level unobservables that may affect the types of herbs being used on a patent. The coefficient β_1 captures the change in probability of introducing an herb (and its familiarity) associated with the inclusion of a new inventor, and the coefficient β_2 measures a similar change when the new inventor is a Chinese or Indian. In particular, β_2 captures the incremental effect of a new inventor being of Chinese/Indian descent on the familiarity of the herb.

We present our results of estimating equation (6) in Table A4. In columns (1-2), our dependent variable is the likelihood of introducing a new herb. Compared to the baseline probability of introducing a new herb (17.41 percent), we see that patents with new inventors are associated with a 15 percentage point increase in the probability of introducing a new patent (column 1, t-statistic = 3.57). Furthermore, in Column 2, an addition of a new Chinese/Indian inventor is associated with another 15.3 percentage point increase in the likelihood of introducing a new herb (p-value 0.041). Columns 3-4 show whether the herbs being introduced by these inventors is more or less familiar. In our sample, herbs have an average log frequency of -16.63 with a standard deviation of 2.369. Again, we see that herbs used by new Chinese/Indian inventors are 2.28 log frequencies less similar in addition to the 2.24 by new inventors, almost two standard deviations less familiar than the average herb. An equivalent change in terms of herb names would be from basil (-11.897) to poria (a type of Chinese mushroom, -16.635), or from buckwheat (-14.257) to Picrorhiza (a Himalayan herb, -19.070). While this relationship is by no means causal, there is indeed a strong association between new inventors, especially new ethnic inventors, with filing new and unfamiliar herbs.

Table A4. New Chinese/Indian inventors use less familiar herbs

	(1)	(2)	(3)	(4)
Dep Var:	Pr(Introduce new herb)		Log frequency of new herb	
New inventor	0.150	0.122	-2.666	-2.244
	(0.042)	(0.044)	(0.737)	(0.776)
New ethnic inventor		0.153		-2.286
		(0.075)		(1.147)
Controls				
Number of claims	0.004	0.004	-0.051	-0.051
	(0.002)	(0.002)	(0.029)	(0.029)
Has Chinese/Indian	0.037	-0.041	-0.720	0.455
	(0.120)	(0.111)	(1.985)	(1.824)
Inventor count	0.027	0.028	-0.263	-0.274
	(0.012)	(0.011)	(0.158)	(0.147)
Chinese/Indian count	-0.030	-0.050	0.479	0.785
	(0.078)	(0.077)	(1.277)	(1.278)
Number of herbs	0.000	0.000	0.014	0.013
	(0.003)	(0.003)	(0.026)	(0.026)
Constant	0.023	0.151	-1.255	-3.174
	(0.095)	(0.116)	(1.701)	(1.999)
Assignee FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	758	758	758	758
Adjusted R^2	0.162	0.170	0.148	0.155

Note: Results of estimating equation (6) using OLS. Standard errors in parentheses, clustered at the assignee level.

Section 4. Firm implications

We next provide suggestive evidence regarding the value of herbal patents by first generation ethnic migrant inventors (as measured by citations). In our context, the value will depend on the quality of the knowledge that was previously locked in the ethnic migrant inventor's home region, and on the host country's firms' ability to codify this knowledge. Thus, given similar access to ethnic migrant inventors, firms that are quicker to codify the new knowledge ethnic migrants carry will benefit more. Ideally, we would test this by randomizing firms' abilities to extract and codify new knowledge by ethnic migrants. Since we are not able to do so, we provide correlational evidence that among firms hiring new ethnic inventors (capped firms), firms that are quicker to patent new herbal knowledge will accrue more citations. We test this by comparing the number of citations to patents with new herbs by capped firms to those without.

Below, we present a table showing that patents filed by capped firms during the visa shock period accrue more citations if they include new herbs. We see that patents with new herbs filed by capped firms during the shock period had more citations than any other group. This suggests that when firms have expanded access to ethnic migrant inventors (capped firms during the visa shock period), they are able to generate more valuable patents by introducing new herbs.

Table A5. Average citations per patents with new herbs and without

	Patent has no new herb		Patent has new herb	
	Non-shock period	Shock period	Non-shock period	Shock period
Exempt	8.89	12.05	5.67	7.00
Capped	14.27	15.98	8.99	16.31

Note: Shock period denotes years between 1999-2003.

We formalize this notion, and estimate the following regression using ordinary least squares:

$$\log(1 + Citation\ Count)_{ijt} = \delta Capped_j \times Shock_t \times NewHerb_i + \beta_1 NewHerb_i + \gamma_1 Capped_j \times Shock_t + \gamma_2 Capped_j \times NewHerb_i + \gamma_3 Shock_t \times NewHerb_i + \lambda_t + \phi_j + \varepsilon_{ijt}$$

The coefficient of interest is δ, which denotes the marginal percent increase in citations when a new herb is included on an herbal patent, when the visa cap is relaxed. We include all interactions between our three indicator variables, *Capped*, *Shock*, and *NewHerb*. We control for year fixed effects and assignee fixed effects, and hence *Capped* and *Shock* are dropped. We present estimation results below.

We see that having a new herb on a patent is correlated with a 91 percent increase in citation counts. This suggests that within patents filed by capped firms (who could hire more ethnic migrant inventors), the ones that introduced new herbs turned out to be more valuable. We are careful not to attach a causal interpretation to this measure. For instance, inventor ability may be driving both citation counts and using new herbs, and our results are reflecting better screening abilities of firms.

Table A6. Effect of visa shock on herbal patent citations

	(1)	(2)	(3)
Model:	OLS		Poisson
Dependent Variable:	Log(1+citations)	Log(1+Citations)	Citations
Capped x Shock x NewHerb	0.617 (0.379)	0.651 (0.365)	0.638 (0.338)
NewHerb x Post	-0.166 (0.310)	-0.190 (0.301)	-0.039 (0.208)
NewHerb x Capped	-0.513 (0.200)	-0.495 (0.187)	-0.597 (0.346)
Capped x Post	-0.313 (0.334)	-0.349 (0.336)	-1.035 (0.734)
Constant	1.884 (0.293)	1.782 (0.304)	
Year FE	Y	Y	Y
Assignee FE	Y	Y	Y
Controls	N	Y	Y
Observations	758	758	497
Adjusted R^2	0.127	0.126	-

Standard errors in parentheses, clustered at the assignee level.

Section 5. Alternate specifications

In this section, we present further results using our difference in difference setting. First, we show that our results are robust to relaxing the assumptions we made regarding the firm founding dates, and allowing firms to have delays between firm founding dates and the initial patent application dates. Second, we show that our results are robust to using nonlinear count models, specifically negative binomial and Poisson models. Third, we show that the visa shock has a causal impact on introducing new Chinese/Indian inventors to capped firms. Finally, we show that patents with new Chinese/Indian inventors are more likely to introduce new herbs that are unfamiliar in the US context.

5.1 Relaxing assumptions regarding firm founding year

In our baseline results, we made an assumption on the relationship between the first time an assignee files a patent, and its founding date. Specifically, we try to collect founding dates from Capital IQ for our assignees. When this method fails, we use the earliest application date for that assignee, for all granted USPTO patents as the assignee's founding year. This assumes that a firm is founded upon patent filing. However, in general, the initial patent application date and firm founding date are likely to be different. Most startups do not have patents upon founding, and in the case of very old companies (founded before 1977, when the patent data starts), there may be significant lags. This section presents results relaxing this assumption, and re-validating our results.

We proceed by varying the assumption on the time it takes from the organization founding date to filing the first patent. We first experiment with a 3 year difference, which can be a reasonable upper bound for most start-ups. We also experiment with 6, 9, and 12 years, in case we have older private firms that are not captured by Capital IQ or the USPTO database. We do not alter the founding dates of firms that we are able to obtain data from Capital IQ. We are only adjusting the founding dates of those assignees that we must impute the founding date from the initial patent application date.

Table A7. Relaxing firm founding date assumption and linear models

	(1)	(2)	(3)	(4)
	Dependent Variable: Log patent count (OLS)			
Time to patent:	3 years	6 years	9 years	12 years
Capped x Shock	0.040 (0.020)	0.041 (0.016)	0.042 (0.014)	0.042 (0.013)
Time FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Controls	Y	Y	Y	Y
Observations	9239	9475	9698	9911
Adjusted R^2	0.068	0.068	0.070	0.072
ll				

Standard errors in parentheses, clustered at the firm level.

We see from Table A7 that regardless of the lags between firm founding and patent application date, we see similar results. Across all columns, the coefficients do not vary significantly. In all specifications, we control for Assignee fixed effects, time fixed effects, and other controls. In this section, we present results using the dataset with significant lags.

5.2 Nonlinear results

Our outcome variable is the number of patent counts at the firm-year level. Since this is a count variable, its values are always nonnegative, and have discrete differences. In these situations, some assumptions of OLS may not be met. The literature tends to either take logs of the count variables, or use nonlinear estimation models to overcome such drawbacks. In our baseline, we opted for the first method and took the log number of patent counts plus one, for two benefits.

First, the interpretation of coefficients is more intuitive. Second, nonlinear models in difference-in-difference settings are complicated, and while they produce coefficients that have the same sign as the true difference in difference effect, more generally the signs do not need to be consistent for interaction terms.

In this section, we present results using nonlinear count models. We focus on two models: the fixed-effect negative binomial model, and the quasi-maximum likelihood (QML) estimates based on the fixed-effect Poisson model. Our data is over-dispersed (i.e., variance is greater than the mean), and thus is more suited to estimation through the negative binomial model. We do not use a zero-inflated model because our data does not suffer from zero-inflation, as discussed in the text[75]. The alternative method has the benefit that even if the underlying model is incorrectly specified, the standard errors are consistent. Furthermore, QML standard errors are robust to arbitrary serial correlation patterns, and are robust to concerns of underestimated standard errors common in difference-in-difference settings. We present our results from estimating the two models below.

Table A8. Nonlinear model results

Dependent Variable: Patent Counts	(1) Fixed Effects Negative Binomial	(2) Fixed Effects Poisson (QML s.e.)
Capped x Shock	0.477 (0.230)	0.445 (0.264)
Constant	-2.517	

[75] We find no evidence for an excessive amount of zeros. We follow Cameron and Trivedi (2010) and compare the predicted probabilities of zeros in a Poisson distribution to the observed probabilities. We see a slightly higher probability of zeros in our observations than the Poisson model would predict (0.0089), but we see no difference in predicted counts of zeros using Stata's countfit command, and the contribution of zeros to the Pearson Chi-Square statistic is 0.001, further showing our data does not suffer from over-inflation of zeros.

	(1.027)	
Time FE	Y	Y
Assignee FE	Y	Y
Controls	N	Y
Observations	9911	9911
ll	-1641.332	-1664.330

Standard errors in parentheses. We relax assumptions on firm founding dates as in 5.1. Column (1) uses the negative binomial model to estimate coefficients, and column (2) uses a Poisson model with quasi-maximum likelihood estimates of the standard errors.

We see that the number of patents increased during the visa shock period. We do not include controls in the specification for column (1) because of convergence issues. For all models, the coefficient is positive, suggesting a positive effect of the visa shock on patenting. For columns (1), the effect size is significant with p=0.038, and for column (2), p=0.092.

5.3 Triple differences

In a previous version of this paper, we estimated a triple differences model for the impact of the visa shock on the likelihood of observing a Chinese/Indian inventor. First, for each herbal patent, we obtained a non-herbal (control) patent that has the same clinical use as the focal herbal patent. We measure whether we are more likely to see Chinese/Indian inventors on herbal patents than on non-herbal patents. This matching captures the notion that Chinese/Indian inventors are more likely to bring knowledge about herbs to the US. We found that it is more likely that Chinese/Indian inventors file herbal patents non-herbal patents that have similar clinical use.

We incorporate our difference-in-difference setting to this result. In the previous paragraph, we hinted that Chinese/Indian inventors file herbal patents at a greater rate than non-Chinese/Indian inventors. The difference-in-difference setting allows us to see whether this effect is being driven by first generation Chinese/Indian invenors. Assuming the visa shock increased the number of first generation migrant ethnic inventors, we see whether the increase in ethnic inventors increases the rate at which herbal knowledge is codified. This allows us to specify a triple difference model, with all pairwise interactions between 1) control and herbal patents, 2) capped assignees and exempt assignees, and 3) during the visa shock and not. The triple interaction term here measures the increase in the rate of codification of herbal knowledge, caused by first generation migrant inventors.

We present the triple-difference results below.

Table A9. Triple Difference Estimates

	(1)	(2)	(3)	(4)
	Fraction	Has Ethnic	Fraction	Has European
	Ethnic		European	
Herbal Patent (HP)	0.0948^{**}	0.165^{**}	-0.0934^{**}	-0.00674
	(0.0442)	(0.0816)	(0.0461)	(0.0340)
Cap-subject assignee	0.00545	-0.0356	-0.0865^{***}	-0.0955^{***}
(CS)	(0.0224)	(0.0420)	(0.0280)	(0.0225)
Shock	0.00190	0.0264	-0.0820^{**}	-0.0400
	(0.0303)	(0.0639)	(0.0412)	(0.0348)
Herbal x Shock	-0.0410	-0.145	0.0894	-0.0431
	(0.0619)	(0.107)	(0.0696)	(0.0677)
Capped x Shock	-0.0133	-0.0431	0.0972^{**}	0.0508
	(0.0327)	(0.0664)	(0.0444)	(0.0381)
Herbal x Capped	-0.0746	-0.145^{*}	0.0346	-0.0299
	(0.0461)	(0.0836)	(0.0495)	(0.0379)
DDD	0.116^{*}	0.247^{**}	-0.146^{*}	0.0117
	(0.0665)	(0.113)	(0.0747)	(0.0715)
Time Trend	0.00533^{***}	0.00823^{***}	-0.000279	0.00100
	(0.000744)	(0.00117)	(0.00138)	(0.00136)
Citations Count	-0.000166	0.000176	0.00149^{***}	0.00130^{***}
	(0.000179)	(0.000304)	(0.000268)	(0.000217)
Inventor Count	0.0215^{***}	0.0629^{***}	-0.0362^{***}	0.0106^{***}
	(0.00570)	(0.00632)	(0.00401)	(0.00292)
Constant	-10.62^{***}	-16.41^{***}	1.499	-1.091
	(1.489)	(2.347)	(2.771)	(2.712)
Observations	4148	4148	4148	4148
Adjusted R^2	0.062	0.138	0.079	0.020

This table presents estimation of equation (1) using ordinary least squares. The dependent variables are the fraction of or an indicator for Chinese/Indian inventors for a given patent (columns (1)-(2)), and the fraction of or an indicator for European inventors (columns (3)-(4)). Herbal and Capped are indicators for whether a patent is an herbal patent or whether the assignee is subject to the H-1B visa cap. Shock is an indicator for when the visa cap was increased (years 2000-2005). Standard errors are clustered at the assignee (employer) level. Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We see that the fraction of herbal patents with Chinese/Indian inventors increased significantly, suggesting that the rate of codification of herbal knowledge is greater for first generation migrant inventors. This suggests that migrant inventors enter the host country and bring knowledge that is previously not in the host context.

5.4 Permutation test

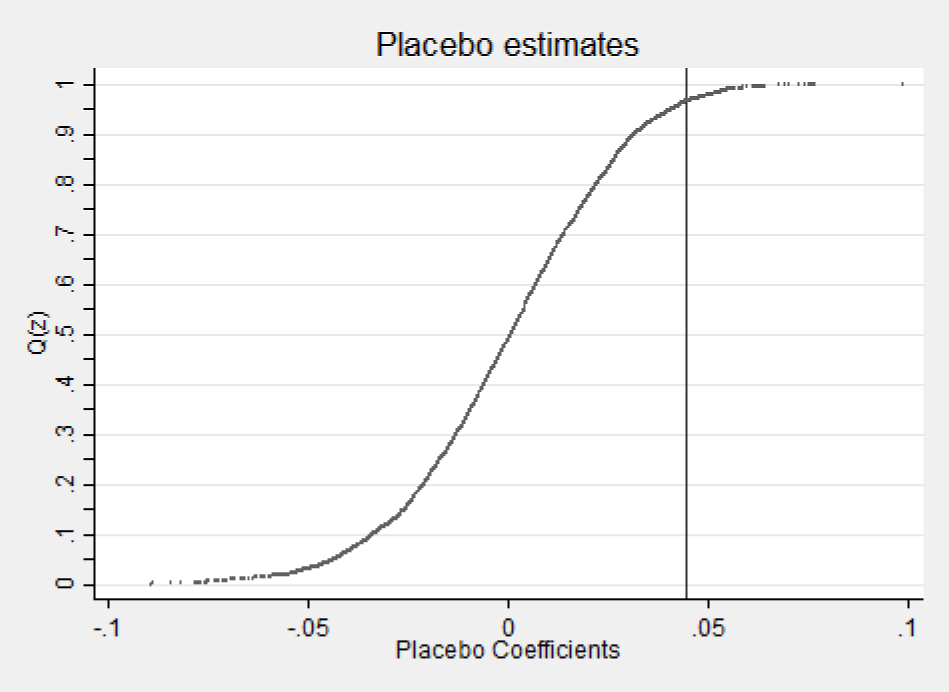


Figure A4. *Cumulative distribution of coefficients from placebo test.* From our sample of 401 herbal patent assignees, we randomly select 73 assignees to be exempt from the visa cap, and also randomly select a consecutive 5-year period to be our placebo H1B visa shock period, and run specification (1) as above, saving the coefficient on the difference in differences (DD) estimate. We repeat this process for 3,200 random treated firm-treatment window pairs. We select random placebo pairs based on two dimensions – assignment of the visa cap to assignees (done 100 times each), assignment of a visa shock period (done 32 times each for the 32 different possible 5-year time periods) for a total of 100×32 = 3,200 random placebo pairs. We plot the cumulative distribution function of the resulting DD coefficients (Q ̂(δ)) in Figure A4. Similar to a p-value, if the visa shock positively affected herbal patenting behavior, we would expect our coefficient to be significantly larger than random (high Q ̂(δ)), and thus appear near the upper right tail of the cumulative distribution function. We reject the null hypothesis of zero effect of the visa shock on herbal patenting if 1-Q ̂(δ)>0.05. The permutation test does not make assumptions about the error structure, and thus is robust to concerns of serial correlation.

Section 6. Summary statistics

In this section, we provide alternative summary statistics on our herbal patents, before and after the visa shock. First, we report t-test results of patent-level variables in our dataset across patents by capped assignees and patents by non-capped assignees, for all time periods. Next, we again present summary statistics on patent characteristics across capped/non-capped assignees, but for the pre-treatment period. This second table allows us to look at any systematic differences in the patents by capped assignees and non-capped assignees.

6.1 Summary statistics across the entire sample

Table A10. t-Test Results

	Capped firms		Exempt firms		Difference	
	mean	sd	mean	sd	b	t
Count of Chinese/Indian inventors	0.31	0.68	0.51	0.99	0.21	(2.19)
Recombined	0.31	0.46	0.29	0.46	-0.02	(-0.39)
Fraction of inventors Chinese/Indian	0.10	0.23	0.16	0.27	0.05	(2.04)
Inventors hired (All)	1.82	1.91	2.15	1.70	0.34	(1.98)
Inventors hired (Chinese/Indian)	0.22	0.59	0.35	0.89	0.13	(1.57)
Application year	2001.60	5.54	2000.17	6.57	-1.43	(-2.27)
Number of claims	15.94	13.41	14.80	10.50	-1.13	(-1.04)
Number of inventors	2.66	1.93	2.77	1.57	0.12	(0.73)
Firm age	9.74	10.12	17.70	9.43	7.96	(8.47)
Firm founding year	1991.87	9.86	1982.47	7.80	-9.39	(-11.68)
New herbs used	0.35	2.44	0.33	0.75	-0.02	(-0.15)
Herbs used	6.39	13.39	6.57	10.40	0.18	(0.17)
1(New herb used)	0.08	0.24	0.14	0.33	0.06	(1.84)
Log Ngram of least familiar herb	-16.38	2.65	-16.19	2.70	0.19	(0.71)
Log Ngram of most familiar herb	-13.47	2.71	-13.41	3.02	0.05	(0.18)
Log Ngram of median herb	-14.90	2.26	-14.74	2.46	0.16	(0.67)
Observations	635		123		758	

We see that patents by capped assignees are more likely to have, and also more likely to have more Chinese/Indian inventors. Capped firms are also more likely to hire inventors. Capped firms tend to file patents later on, but exempt firms tend to be older when they file patents. Finally, exempt firms are more likely to use new herbs.

6.2 Summary statistics before the visa shock

Table A11. Herbal patent characteristics before visa shock (pre 1999)

	(1)		(2)		(3)	
	Capped		Exempt		Difference	
	mean	sd	mean	sd	b	t
Chinese/Indian inventor count	0.23	0.59	0.28	0.49	0.05	(0.68)
Recombined	0.25	0.44	0.25	0.43	-0.01	(-0.13)
Fraction Chinese/Indian	0.08	0.20	0.13	0.27	0.05	(1.44)
Inventors hired	1.94	1.65	2.07	1.50	0.13	(0.55)
Inventors hired (Chinese/Indian)	0.21	0.58	0.18	0.43	-0.03	(-0.44)
Application year	1995.33	4.32	1994.72	4.61	-0.61	(-0.89)
Claims	17.42	15.06	16.95	9.50	-0.47	(-0.28)
Inventor count	2.33	1.56	2.46	1.38	0.12	(0.57)
Firm age	7.34	7.67	13.53	7.53	6.19	(5.42)
Firm founding year	1987.99	8.52	1981.19	6.73	-6.80	(-6.26)
New herbs used	0.74	4.37	0.54	0.96	-0.19	(-0.56)
Herbs used	5.78	10.83	2.81	2.97	-2.98	(-3.38)
1(New herb used)	0.13	0.30	0.22	0.37	0.09	(1.62)
Log Ngram of least familiar herb	-15.81	2.25	-15.97	2.83	-0.16	(-0.38)
Log Ngram of most familiar herb	-13.14	2.34	-13.88	3.09	-0.74	(-1.64)
Log Ngram of median herb	-14.39	1.88	-14.88	2.66	-0.49	(-1.28)
Observations	189		57		246	

Both groups have a similar number of inventors and claims. Both groups' patents are equally likely to contain synthetic compounds, as measured by our *Is Synthetic* variable, and use herbs that are equally frequent in the English language. Cap exempt assignees' patents in our sample are slightly older on average, and are generally more likely to have inventors with Chinese/Indian names.

6.3 Alternate graphs using Stata's binscatter command

In this section, we plot the number of patents per assignee-year for capped and exempt firms over time. We also plot the recombination probabilities over time for any given herb, and the recombination probabilities by a specific ethnicity.

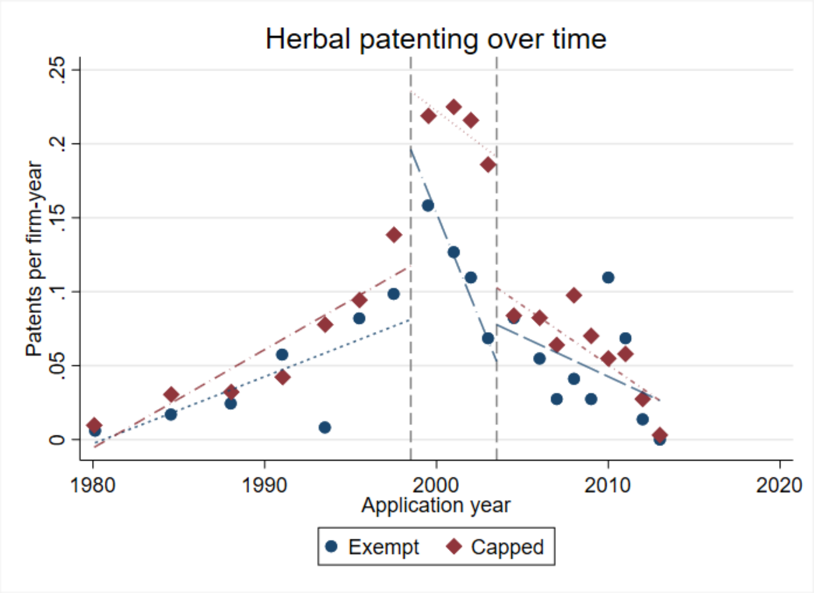


Figure A5. Binscatter plot of patenting over time

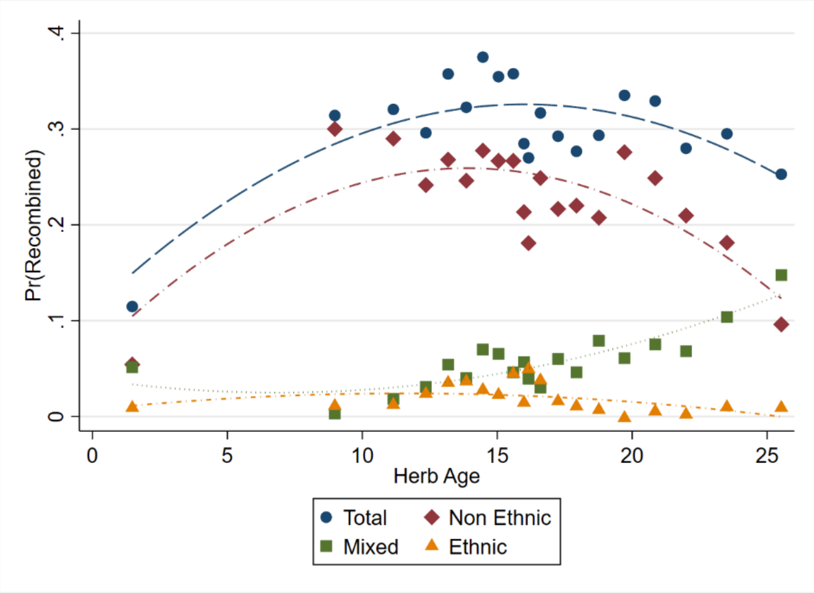


Figure A5. Binscatter plot of recombination probabilities over time by ethnic groups

Section 7. Coding recombination, ethnicities, and assignees

7.1 Coding the Recombined variable

We create a variable is_recombined to indicate patent applications comprising herbs combined with synthetic compounds.To do this categorization, we used the Derwent classification for each patent. The Derwent Patent class is a manually curated standardized classification system for patents maintained by Thomson Reuters and the classification is more industry centric than technology centric. After analyzing various mixes of Derwent classes in herbal medicine patent records, we concluded that herbal medicine patent records containing Derwent classes B05, B06, or B07 comprised a mix of herbal medicines and other synthetic compounds/drugs. Inthe absence of any of these classes, the composition is purely made up of herbal medicines. In Derwent classification, the B class refers to 'pharmaceuticals.' Subclass B05 refers to 'other organics,' (B05 other organics -aromatics, aliphatic, organo-metallics, compounds); B06 to 'inorganics,' (inorganics - including fluorides for toothpastes etc.) and B07 to 'general'(tablets, dispensers, catheters (excluding drainage and angioplasty), encapsulation etc.) B04 refers to 'natural products and polymers,' which also includes herbal medicine patents but does not contain synthetic compounds. B05, B06, and B07 are the only three classes in B (pharmaceuticals) that contain synthetic Western drugs. Thus, a presence of these three classes signifies a combination of synthetic compounds/drugs with herb. Fifty random abstracts of patent records having any of these three classes and 50 random abstracts of patent records with absence of all of these three classes were studied to confirm the effectiveness of using Derwent classes to code the 'Recombined' variable and this result was independently verified by two different coders and checked by the researchers.

7.2 Coding ethnicities

Probabilistically, surnames such as Xing are more likely to be associated with Chinese individuals than with other ethnicities. We build on this insight and utilized an open-source name categorizer “ethnicityguesser” to categorize inventors’ ethnicities. The software is based on the Natural Language ToolKit (NLTK) package in Python, and it comes pre-packaged with a set of names and associated ethnicities. As a robustness check, we compare our ethnicity classification results when using different training sets and against Ambekar et al. (2009) who use state of the art hidden Markov models and decision trees for classification.

Table A12. Correlations across Ethnicity measures

	Asian1 (surname)	Asian2 (surname)	Asian1 (full)	Asian2 (full)
Chinese/Indian1 (first)	1			
Chinese/Indian2 (first)	0.9637	1		
Chinese/Indian1 (full)	0.9503	0.9195	1	
Chinese/Indian2 (full)	0.9297	0.9394	0.963	1

Note: Comparison of the two training sets provided by kitofans’ ethnicityguesser.As expected, we see that for all classifications, there is a high correlation across the two measures, whether we use the full name or just the surname as the tokens.

We also compare our results to the Name Ethnicity Classifier created by Ambekar et al. (2009). If we have a high correlation between our measure of Chinese, Indian and European with the Name Ethnicity Classifier’s Asian and Greater European categories, we would be confident about our measures of ethnicity. We randomly sample 10% (1,219) of our inventors’ names and submit this to the Name Ethnicity Classifier’s website. We present the results below. We see that 94 percent of our Chinese inventors are categorized as Asian, and 90 percent of our Indian inventors are categorized as Asian. Generally, our classification of European coincides with the categorization of Europeans by Ambekar et al (2009). Overall, the results reflect positively on our classification of ethnicities.

Table A13. Comparison of ethnicityguesser performance to benchmark ethnicity classification product

kitofans	Asian	GreaterAfrican	GreaterEuropean
african	5	2	0
arabic	0	0	2
chinese	115	0	7

czech	16	5	28
danish	1	0	25
french	12	7	170
german	1	0	54
greek	4	1	11
indian	70	4	3
irish	0	0	31
italian	7	2	21
japanese	133	3	2
jewish	14	11	163
korean	63	1	6
muslim	6	14	2
portugese	4	1	12
russian	0	0	3
slavic	0	0	7
spanish	7	5	51
swedish	3	1	43
swiss	2	2	36
ukranian	1	1	10
vietnamese	5	0	3

Note: Comparison of classification results using kitofans' ethnicityguesser and Ambekar et al (2009). We read the table as follows. Of the 122 names classified as Chinese using ethnicityguesser, 115 are classified as Asian in Ambekar et al (2009).

7.3 Educational background of Chinese/Indian inventors

Inventor backgrounds can also provide information about whether herbal patent inventors are more likely to be first generation migrant inventors. Our 3,182 herbal patent grants (U.S. and foreign assignees) contains 6,119 unique inventors, of which 1,208 unique inventors are of Chinese or Indian ethnicities. We randomly sample 552 inventors from the Chinese/Indian inventor population (45% of unique Chinese/Indian inventors in herbal patents sample) and attempt to search for their educational history in LinkedIn. To do so, we search for individuals in LinkedIn using the inventor's and assignee's names. If there is a profile that 1) has a match on the inventor name, 2) match for the assignee of interest 3) near the time period the patent application was submitted, we code this as a successful search. We successfully found 84 profiles on LinkedIn (15% of Chinese/Indian inventors that we looked up on LinkedIn), but we drop 20 individuals who do not list their educational details. For each Chinese/Indian inventor left, we document the educational background of the individuals. We document whether the inventor was educated solely in India, U.S., or China, or whether they were educated elsewhere and moved to the U.S. Of the sample, about one third of the individuals were educated solely in India and the U.S. each. About 20% of individuals were educated first in China, then moved to the U.S. The remaining inventors were educated just in China (9%) or educated in India, then educated in the US (3%). In summary, a disproportionate fraction of matched Chinese/Indian inventors filing herbal patents who we looked up on LinkedIn, were educated in China/India, indicating that they were indeed first generation migrant inventors.

Table A14. Educational background of Chinese/Indian inventors

Educational Background	Count	Percentage
India	22	34.38%
US	20	31.25%
China to US	14	21.88%
China	6	9.38%
India to US	2	3.13%
Total	64	100%

We also look at whether herbal patents have more inventors that were educated in China/India. Herbal patents are much more likely to have inventors educated solely in India, and similarly for Chinese educated individuals. On the other hand, inventors educated abroad who moved to the US are less likely to write herbal patents. Inventors educated solely in the US are less likely to write herbal patents, despite their being ethnically Indian/Chinese.

Table A15. Educational background of Chinese/Indian inventors by patent type

Educational Background	Control Patent	Herbal Patent
India	5	17
US	14	6
China to US	9	5
China	3	3

India to US	2	0
Total	33	31

Finally, we see whether the visa shock increased the number of foreigners writing patents. Towards this, we look at whether patents written during the visa cap increase have more inventors that were educated outside the US. The shock seems to have increased the proportion of Indian inventors, but decreased all other types of inventors.

Table A16. Educational background of Chinese/Indian inventors over time

Educational Background	Non-Shock	Shock
China	3 (9.09%)	3 (9.68%)
China to US	9 (27.27%)	5 (16.13%)
India	5 (15.15%)	17 (54.84%)
India to US	2 (6.06%)	0
US	14 (42.42%)	6 (19.35%)
Total	33	31

7.4 Assignee classification

We categorized assignees into three broad groups: individuals, U.S. based firms and Universities, and Foreign firms and Universities. We merge our herbal patents dataset with two external datasets for this process: USPTO's PatentsView database, and Capital IQ. PatentsView contains disambiguated assignee data, and classifies each assignee into U.S. Company or Corporation, Foreign Company or Corporation, Individuals, and so forth. Capital IQ provides researchers with corporate headquarters for a company. In our sample of U.S. assignees, there are 401 unique assignees, 73 of which are cap exempt.

In our original dataset of the universe of herbal patents, there were a total of 7,157 patents. We obtain the geographical data for 3,183 patents from PatentsView, 3,562 from Capital IQ and 412 from manual searches. We find that 2,512 of our patents are filed by individuals, 2,851 by foreign companies, and 1,794 by firms/Universities based in the US. We also obtained a list of H1B visa cap-exempt employers from a 3rd party online employment entity[76]. The online list contains 12,479 employers who have been categorized as exempt from the H1B visa cap. We matched these employers to our list of assignees, and further searched for "university" and "college" to construct a list of assignees that are exempt from the H1B visa cap (*CAP*). Out of the 4,179 total number of unique assignees in our herbal patent sample, 158 unique assignees are exempt from the H1B visa cap. In our sample of U.S. assignees, there are 401 unique assignees, 73 of which are cap exempt.

7.5 Sample restrictions

Our sample restrictions are based on three considerations: 1) location of firm, 2) founding dates/final patenting dates, 3) patent grants. We first explain the data sources, then delineate the sample restriction process and rationale, and finally discuss the magnitudes of these changes. We use assignee locations from PatentsView when available, and also use headquarter locations provided by Capital IQ. We also use Capital IQ to obtain founding dates when available, and impute founding dates based on patent application years. We also obtain the date of the last patent filed by an assignee. Application status (granted or not) is obtained through PatentsView.

In the full dataset, there are 4,179 assignees, 1,368 of which are granted any patents. Restricting the headquarter location to US based assignees brings this down to 1,037 firms, 585 of which are granted any patents. Finally, we restrict the sample to firms founded before 2004 that continued filing patents through 2000. We thus have 633 firms found before 2004 that filed at least one patent after 2000, 401 of which are granted any patents.

[76] Source: http://www.myvisajobs.com/Search_Visa_Sponsor.aspx

The largest change is in excluding the non-US firms, but our identification strategy does not specify in which direction foreign firms' patenting should move. The restriction on founding dates and last patenting dates is necessary because patenting outside of our visa shock period may bias the results downwards. Finally, the restriction on firms with any granted patents can be relaxed to obtain similar, but noisier results. Using all 633 firms that are filing patents, we observe a 2.29 percent increase in patent filings (p=0.0998) while including assignee fixed effects, application year fixed effects, and time varying firm controls.

Section 8. Herb characteristics

For all granted herbal patents, we collect a list of herbs mentioned in the title and abstract of the patent. For each of these herbs, we collect the Google N-gram score of the herb in the patent application year. Google N-grams provides users with the raw frequency of words in all American English books digitized by Google, for a given year until 2009. We define the Familiarity of the herb to be the log of the N-gram score. Since we have the universe of herbal patents, we can also obtain the minimum year in which an herb was used in a patent. We define the Year Introduced variable as the minimum application date across all patents using a specific herb. We present a select list of herbs, and their characteristics below.

Table A17. Most frequent and least frequent herbs

Herb Name	Patents using herb	Familiarity	Year Introduced	Notes
corn	82	-10.6883	1984	Most frequent herb
soybean	71	-13.1475	1985	
soy	66	-12.1929	1994	
green tea	65	-14.0068	1996	
vegetable oil	60	-13.5257	1984	
		⋮		
rosmarinus	12	-16.8946	1995	

echinacea purpurea	5	-16.8609	1999	Median familiar herb
vitis vinifera	1	-16.8603	2008	Median familiar herb
phytolacca	3	-16.8545	1992	
		⋮		
paullinia cupana	1	-19.2911	1987	
catharanthus roseus	1	-16.6776	1986	
zygophyllaceae	1	-17.6729	1986	
matico	1	-19.8119	1986	
salicornia	1	-16.6955	1985	Least frequent herb

Note: Most and least frequent herbs, including the median familiar herbs. Patents using herb are counts of patent grants by US based assignees (corporations and universities) that were active during 1999-2003. Familiarity is measured as the log of the Ngram of the herb when the patent application was filed; here we report the mean log Ngram across all patents using an herb. Year Introduced is the first year in our sample in which the herb was used.

Section 9. Impact of restrictive immigration policies

In this section, we present results showing that restrictive immigration policies can deter recombinatory innovation.

Table A18. Effect of restrictive migration policies post 2004

	(1)	(2)	(3)
		Dependent variable: log(1+patent count)	
Capped x Post2004	-0.032***	-0.018**	-0.036***
	(0.010)	(0.009)	(0.010)
Capped	0.037***	0.023***	
	(0.006)	(0.006)	
Post2004	-0.005		
	(0.007)		
Constant	0.040***	-0.008	0.015
	(0.004)	(0.009)	(0.010)
Time FE	N	Y	Y
Assignee FE	N	N	Y
Observations	8998	8998	8998
Adjusted R^2	0.008	0.036	0.038

Cluster robust standard errors in parentheses, clustered at the assignee level. Observations are at the assignee-year level, for all years an assignee (firm or university) was in operation. F=We use the tsfill command in Stata to fill in missing assignee-year pairs. The assignee-year level dataset is thus an unbalanced panel consisting of 8,998 observations (an average of 22.4 years of observations for 401 assignees). The dependent variable is the log of the number of herbal patents filed by an assignee in a given year. Capped is an indicator for whether the assignee is subject to the visa cap; Post2004 is an indicator for years 2004 and onwards. Percentage changes are calculated as $100 \cdot \left(e^{\beta} - 1\right)$.

$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Section 10. Recombination by non-ethnic teams

It is worth noting we observe non-ethnic (no ethnic inventor) teams participating in recombination as well as mixed teams. This is puzzling because we hypothesized that recombination requires knowledge of both the local context and the foreign context, suggesting recombination should be driven by mixed teams. We argue indirect spillovers from prior collaborations and prior herbal patent codification can allow non-ethnic teams to recombine. We illustrate these mechanisms with empirical evidence.

10.1 Spillovers through co-inventors

We first test the relationship between prior ethnic exposure and recombination probabilities. Specifically, we test whether the probability of recombination varies with prior exposure to ethnic inventors, and whether that rate varies across the different types of teams. We discuss the graphical results first.

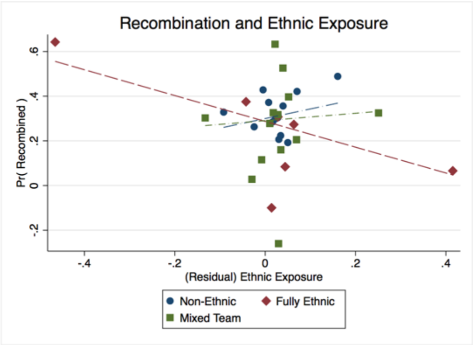


Figure A6. Recombination probabilities and prior ethnic co-inventor exposure

We see that for non-ethnic and mixed teams, the probability increases with ethnic exposure, while fully ethnic teams exhibit a downward slope. We test this formally by estimating the following equation:

$$Recombined_{ijt} = \beta_1 Exposure + \beta_2 NonEthnic + \gamma Exposure \times NonEthnic + \lambda_t + \phi_j + \epsilon_{ijt}$$

where *Exposure* is the average prior exposure to ethnic inventors across all inventors in a patent, and *NonEthnic* is an indicator variable for patents with no ethnic inventors. The main coefficient of interest is γ which captures the differential increase in likelihood of recombination from an increase in ethnic exposure for non-ethnic teams.

Table A19. Increase in prior ethnic collaboration increases recombination

	(1)	(2)
	Dependent Variable: Recombined	
Ethnic Exposure	-0.491^{*}	-0.568^{*}
	(0.260)	(0.300)
Non-Ethnic	-0.020	0.063
	(0.070)	(0.073)
Non-Ethnic x Ethnic Exposure	1.104^{*}	1.067^{*}
	(0.612)	(0.631)
Assignee FE	Yes	Yes
Application Year FE	No	Yes
Controls	No	Yes
Observations	501	454
Adjusted R^2	0.212	0.205

Cluster robust standard errors in parentheses, clustered at the assignee level (* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$).
Exposure is skewed, ranging from 0 to 1 with a mean of 0.028 and a standard deviation of 0.128. The sample is at the patent level, and the sample size differs because of assignee fixed effects: observations without assignee level variation will be dropped.
We see from column 1 that the coefficient on Non-Ethnic x Exposure is positive and significant (p=0.073). This suggests that a one standard deviation increase in Exposure increases recombination probabilities quicker than it does for fully ethnic teams. This pattern is robust to including controls and time fixed effects (column 2)

10.2 Spillovers through prior inventions

Another mechanism by which non-ethnic inventors can recombine knowledge is if the effects of the herb have been codified in the past. Codified knowledge regarding herbs significantly lowers the cost of recombination even to non-ethnic inventors. One consequence of this mechanism would be that the herbs non-ethnic inventors use will be older (and thus more likely to be codified). We perform this indirect test by estimating the following regression equation:

$$\begin{aligned} &MedianHerbYear_{ijt} \\ &\qquad = \beta_1 NonEthnic_i \times Recombined_i + \beta_2 Recombined_i + \beta_3 NonEthnic_i + \lambda_t \\ &\qquad + \phi_j + \epsilon_{ijt} \end{aligned}$$

where MedianHerbYear is the median age of the herb on a patent, NonEthnic and Recombined are nidicators, and we include assignee year fixed effects and firm fixed effects when appropriate. The coefficient of interest is β_1, which compares whether recombined patents use older herbs on average, and whether this is more pronounced for non-ethnic teams.

Table A20. Recombined patents by nonethnic inventors have older herbs

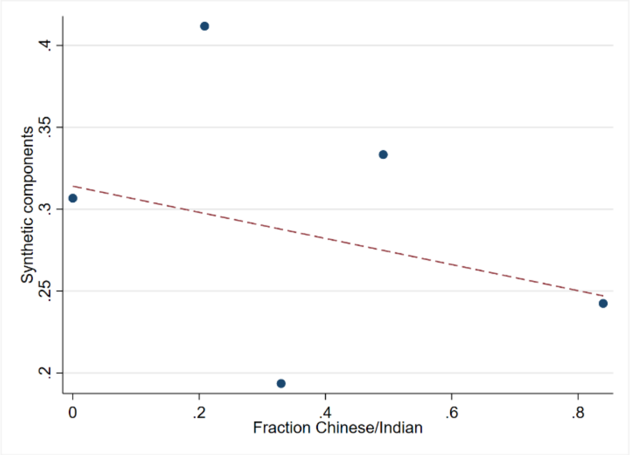
	(1)	(2)	(3)
		Years since median herb was first used	
Non Ethnic	-2.010	-1.107	-1.526
	(1.277)	(1.071)	(1.942)
Recombined	-2.001	-1.961	-2.018
	(1.965)	(1.793)	(1.801)
Non Ethnic x Recombined	4.807	3.503	3.721
	(2.221)	(2.015)	(2.017)
Assignee FE	Y	Y	Y
Application Year FE	N	Y	Y
Controls	N	N	Y
Observations	501	494	494
Adjusted R^2	0.318	0.512	0.522

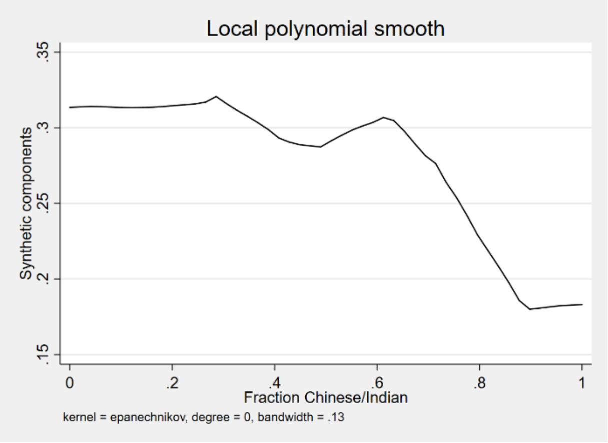
Robust standard errors in parentheses

From column 1, we see that the median herb is younger on non-ethnic teams compared to teams with ethnic inventors, but this number is statistically insignificant at the 10 percent level (p=0.116). Interestingly, the median herb in a Recombination patent is younger than reuse patents, but this effect is also statistically insignificant (p=0.309)[77]. The interaction term is positive and significant (p=0.031), suggesting that the median herb on a recombination patent is older than a reuse patent, but only for the non-ethnic teams. This is consistent with our hypothesis that non-ethnic teams need to build on prior codification of herbal knowledge, while teams with ethnic inventors do not have such restrictions.

Section 11. Alternate tests of recombination

First, to visualize the relationship, we include binned scatterplots and local polynomial smoothing estimates of the relationship between the usage of synthetic compounds and the fraction of ethnic inventors. The graphs are obtained via Stata's *binscatter* and *lpoly* commands, respectively.





[77] Note that this is only the case when controlling for assignee fixed effects. Thus the interpretation is that the median herb in a recombination patent is younger than a reuse patent within assignees that have some ethnic inventors.

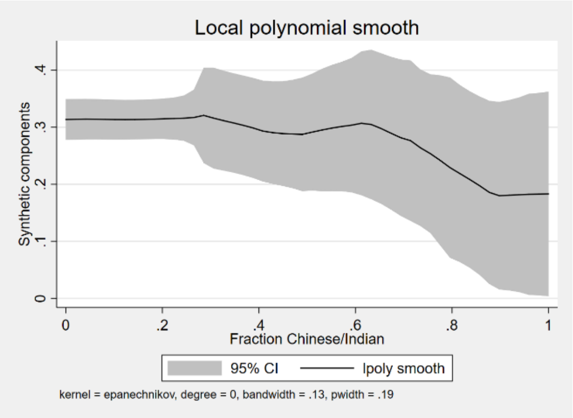


Figure A7. Binned scatterplots and local polynomial smoothing curves of recombination

The graphs suggest that there is a nonlinear relationship between the two variables. For low Chinese/Indian inventor patents, the use of synthetic compounds is relatively high, while the downward slope is primarily being driven by patents with a high fraction of Chinese/Indian inventors. We include a version of the second graph with confidence intervals as well. This last graph shows that while the probability of recombination is decreasing as the fraction of Chinese/Indians increases, the estimates are also increasingly noisier.

We next investigate whether recombination probabilities are affected by inventor counts and ethnic inventor counts independently. We estimate 3 simple models, all with recombination as the main dependent variable, and with inventor counts, ethnic inventor counts, and both inventor and ethnic inventor counts as the independent variables. We present the results below.

Table A21. Recombination and inventor counts

	(1)	(2)	(3)
	Dependent Variable: Recombination		
Inventor count	0.015		0.024
	(0.011)		(0.014)
Ethnic inventor count		-0.037	-0.066
		(0.032)	(0.037)
Assignee FE	Yes	Yes	Yes
Application Year FE	Yes	Yes	Yes

Observations	501	501	501
Adjusted R^2	0.216	0.214	0.218

Cluster robust standard errors in parentheses, at the patent assignee level. Sample at the patent level. Note: since we include assignee fixed effects, assignees with only one patent are dropped, hence the different sample size.

We see that while inventor count is positively associated with recombination, the effect is statistically insignificant (p=0.186). Similarly, Ethnic inventor count is negatively associated with recombination, but again statistically insignificant (p=0.251). Including both as independent variables magnifies the effect of both variables. Furthermore, the effects have greater statistical significance: p=0.097, p=0.082 for inventor count and ethnic inventor count respectively.

Section 12. Tests of inverted-U relationships

Next, we utilize Simonsohn (2017) to formally test whether an inverted U-shaped relationship exists for recombination over time. Intuitively, an inverted U-shape implies the existence of a cutoff x_c such that if $x \leq x_c$, then x and y are positively related, and if $x > x_c$, then x and y are negatively related. Simonsohn (2017) chooses the cutoff based on a "Robin Hood" algorithm (which he shows is robust to errors), estimates an interrupted regression, and checks if the coefficients for high and low values of x (1) have opposite signs and (2) are independently statistically significant according to a pre-specified threshold. We implement this test on the herb-level demeaned variables for recombination and time. That is, for patent i using herb h, we define the demeaned recombination variable $\widetilde{y_{ih}} = y_{ih} - \bar{y}$ and the demeaned time variable $\widetilde{x_{ih}} = x_{ih} - \bar{x}$ where $\bar{y} = \frac{1}{N}\sum_i y_{ih}$, $\bar{x} = \frac{1}{N}\sum_i x_{ih}$. We follow Simonsohn and estimate the cutoff $\widetilde{x_c}$ using the Robin Hood algorithm, and estimate the interrupted regression using a significance threshold of 0.05. We present graphical results below (note the y-axis denotes recombination and the x-axis denotes time since herb introduced).

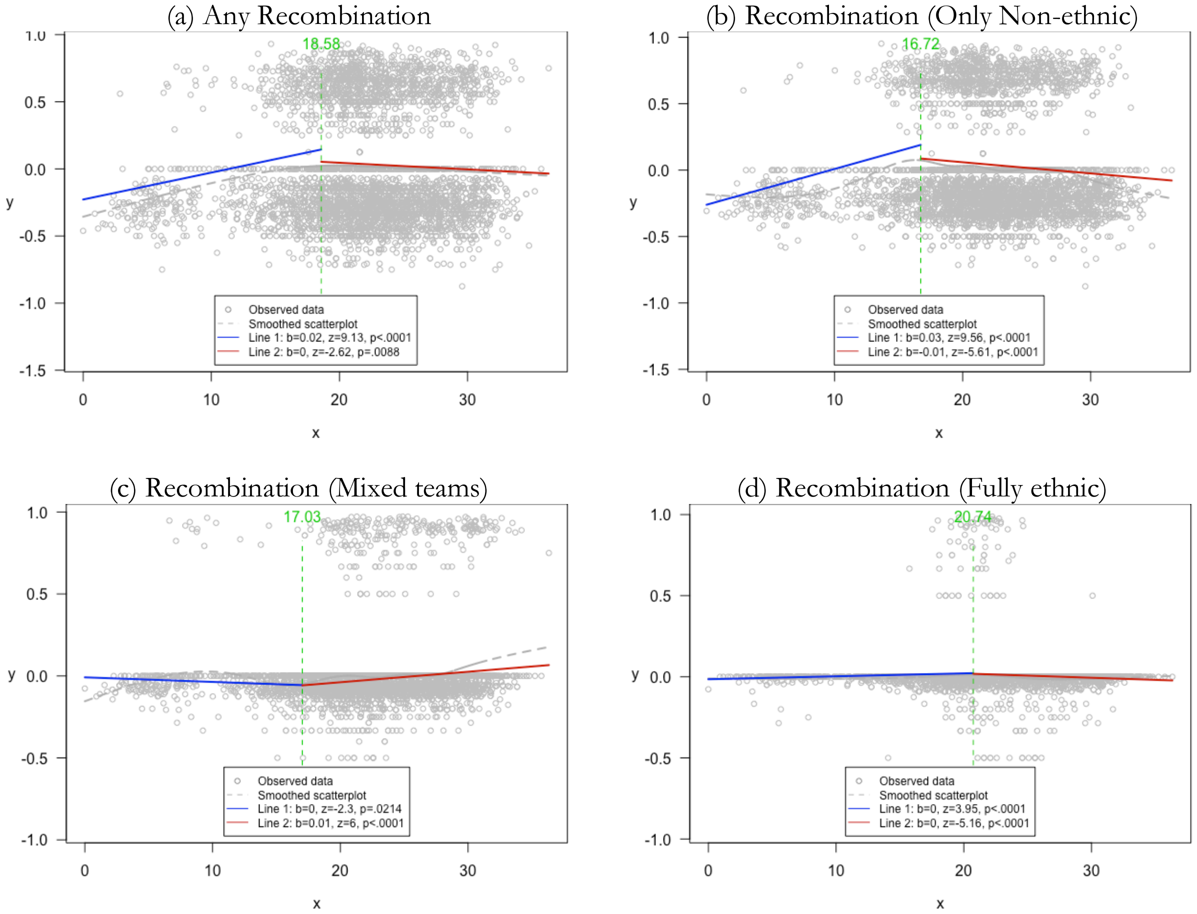


Figure A8. Recombination probabilities over time, by ethnic teams. (Z-scores represent b/se)

As we see from Figure A8, we observe inverted U-shape patterns for some ethnic groups (Non-ethnic and fully ethnic teams), but the opposite for mixed teams. Overall, there is an inverted U-shape relationship with recombination increasing for the first 18.6 years ($\beta_{low}^{overall} = \mathbf{0.020}$, *p*<0.001), but decreases afterwards ($\beta_{high}^{overall} = -\mathbf{0.005}$, *p*=0.009). Interestingly, recombination by entirely non-ethnic teams is increasing for the first 17 years an herb is introduced ($\beta_{low}^{nonethnic} = \mathbf{0.027}$, *p*<0.001), while recombination by mixed teams is decreasing during the same time period ($\beta_{low}^{mixed} = -\mathbf{0.003}$, *p*=0.021). Subsequent recombination decreases for entirely non-ethnic teams ($\beta_{high}^{nonethnic} = -\mathbf{0.008}$, *p*<0.001), while it increases for mixed teams ($\beta_{high}^{mixed} = \mathbf{0.006}$, *p*<0.001). We see similar patterns for fully ethnic teams and entirely non-ethnic teams, but the impacts are smaller ($\beta_{low}^{fullethnic} = \mathbf{0.002}$, *p*<0.001, $\beta_{low}^{fullethnic} = -\mathbf{0.003}$, *p*<0.001).

While not reported here, we find similar results after controlling for patent characteristics such as inventor counts and number of claims.

Qualitatively, a significant portion of initial recombination is driven by non-ethnic teams. Comparing overall recombination and non-ethnic recombination (panels a and b in Figure A8), we see the cutoff point is smaller for non-ethnic teams (16.72 versus 18.58), and the slope is larger for non-ethnic teams (0.027 versus 0.020). This suggests local inventors (non-ethnic teams) recombine knowledge faster and earlier on. The other team compositions do not significantly drive recombination during this period, with the slope for full-ethnic teams and mixed teams at 0.002 and -0.003, an order of magnitude smaller than non-ethnic teams.

Similarly, while overall recombination is decreasing as more time has passed since an herb's initial use, mixed teams continue to recombine in later time periods. Mixed teams increase the rate of recombination after 17.03 years, with recombination increasing at a rate of 0.006 per year, in contrast to the overall negative trend of recombination at -0.005.

Appendix B: Appendix to Chapter 2

Table B.1. Timeline of reverse engineering and legal events

Year	General community	Broadcom	Atheros	Intersil	Texas Instruments	Other activities
1999				- Intersil releases first PRISM II wireless chips[78] (November 9)		
2000	- Communities such as SeattleWireless and Bay Area Wireless User Group (BAWUG) begin forming around potential for wireless internet					
2001			- Atheros releases first 802.11a solution[79] (October 15)		- Texas Instruments releases a 802.11b chipset[80]	
2002		- Broadcom releases forward compatible 802.11g system-on-a-chip[81] (September 10)	- Atheros driver reverse engineering begins[82] (August 6)	- Intersil and Conexant enter memorandum of agreement[83], ceases open source		- Linksys launches WRT54G using Broadcom

[78] https://wikidevi.com/wiki/Intersil

[79] https://wikidevi.com/files/Atheros/specsheets/AR5000.pdf

[80] http://www.ti.com/pdfs/bcg/ti_acx100.pdf

[81] https://web.archive.org/web/20041211212226/https://www.broadcom.com/press/release.php?id=332500

[82] https://web.archive.org/web/20020806172744/http://team.vantronix.net:80/ar5k/

[83] https://www.eetimes.com/document.asp?doc_id=1178338

				support[84] (September 3)		SOC[85] (December)
2003	- Andrew Miklas notifies community of Cisco/Linksys GPL violation[86] (June 7) - Free Software Foundation takes on coordinating role between community and Cisco/Linksys[87] (September 29)	- Broadcom driver reverse-engineering begins[88] (March 21) - Cisco/Linksys source code release begins[89,90] (July 4) - Cisco/Linksys source code release is complete, except drivers[91] (October 14)	- Atheros releases partially open drivers with a hardware abstraction layer (HAL)[92,93] (June 6)		- Texas Instruments driver reverse-engineering begins[94,95] (March 1) - Texas Instruments driver reverse engineering complete[96] (July 26)	- Cisco acquires Linksys[97] (March 20)
2004	- OpenWRT project started[98] (January)			- Intersil prism54 (islsm) reverse engineering begins		

[84] https://hewlettpackard.github.io/wireless-tools/Linux.Wireless.drivers.802.11ag.html#Prism54softmac
[85] https://www.ifixit.com/Device/Linksys_WRT54G
[86] https://lkml.org/lkml/2003/6/7/164
[87] https://lwn.net/Articles/51570/
[88] https://web.archive.org/web/20030813051755/http://linux-bcom4301.sourceforge.net:80/
[89] https://hardware.slashdot.org/story/03/07/06/2121234/linksys-releases-gpled-code-for-wrt54g
[90] https://lwn.net/Articles/53138/
[91] https://lwn.net/Articles/53780/
[92] https://web.archive.org/web/20030714141027/http://sourceforge.net:80/projects/madwifi/
[93] http://madwifi-project.org/wiki/About/History
[94] https://sourceforge.net/projects/acx100/
[95] https://web.archive.org/web/20030418045308/http://acx100.sourceforge.net:80/
[96] https://web.archive.org/web/20030806124743/http://acx100.sourceforge.net:80/
[97] https://www.networkworld.com/article/2340867/network-security/cisco-buys-home-networker-linksys.html
[98] https://wiki.openwrt.org/about/history

	- HyperWRT project started[99] - Sveasoft releases the for-profit Alchemy[100] (September 1)			for softmac[101] (November 19)		
2005	- Brainslayer releases DD-WRT clone of Alchemy[102] (January 22)	- Broadcom driver reverse-engineering ends[103] (December 8)		- Intersil prism54 (islsm) reverse engineering complete[104] (October 25)		- Linksys releases WRT54GL which supports Linux[105] (December 5)
2006	- "Tomato" released as successor to HyperWRT,[106] serves as basis for ASUS routers,[107] and further modified by users[108]	- Reverse engineered Broadcom driver (b43) included in Linux kernel[109,110] (June 17)	- Atheros HAL ported to Linux from prior reverse engineering work[111], despite copyright concerns[112] (February 23)			

[99] https://en.wikipedia.org/wiki/HyperWRT
[100] https://web.archive.org/web/20041018033025/http://www.sveasoft.com:80/
[101] https://web.archive.org/web/20041122031828/http://jbnote.free.fr:80/prism54usb/
[102] https://en.wikipedia.org/wiki/DD-WRT
[103] https://web.archive.org/web/20051208033349/http://bcm43 .berlios.de:80/
[104] https://web.archive.org/web/20051025000223/http://prism54.org:80/
[105] https://lwn.net/Articles/162429/
[106] https://en.wikipedia.org/wiki/Tomato_(firmware)
[107] https://asuswrt.lostrealm.ca/about
[108] https://github.com/RMerl/asuswrt-merlin/wiki/About-Asuswrt
[109] https://lwn.net/Articles/314313/
[110] https://kernelnewbies.org/Linux_2_6_17
[111] https://web.archive.org/web/20060720110743/http://lists.gnumonks.org/pipermail/ath-driver-devel/2006-February/000179.html
[112] http://zgp.org/pipermail/linux-elitists/2005-June/011205.html

2007			- Software Freedom Law Center declares reverse engineered Atheros driver is cleared of copyright infringement claims[113] (September 27)			
2008	- FSF files lawsuit against Cisco Systems for GPL violation[114] (December 11)		- Reverse engineered Atheros driver ath5k included in Linux kernel[115] (April 17) - Atheros releases ath9k, first fully free and open source wireless driver[116] (July 28)			
2009	- FSF and Cisco Systems settle out of court[117] (May 20)					
2010		- Broadcom releases first fully open wireless driver[118] (September 9)				

[113] https://www.softwarefreedom.org/resources/2007/ath5k-code-analysis.html

[114] https://en.wikipedia.org/wiki/Free_Software_Foundation,_Inc._v._Cisco_Systems,_Inc.

[115] https://kernelnewbies.org/Linux_2_6_25

[116] https://www.fsf.org/news/ath9k

[117] https://en.wikipedia.org/wiki/Free_Software_Foundation,_Inc._v._Cisco_Systems,_Inc.

[118] https://lwn.net/Articles/404248/