Knowledge Flows and IP Within and Across Firms – Economics and Machine Learning Approaches

A dissertation presented

by

Mike Horia Mihail Teodorescu

to

The Strategy Unit at 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

April 2018

© 2018 Mike Horia Mihail Teodorescu

All rights reserved

Dissertation Advisors: Tarun Khanna, Mike Horia Mihail Teodorescu
Shane Greenstein

Knowledge Flows and IP Within and Across Firms – Economics and Machine Learning Approaches

Abstract

Knowledge produced in a firm is a source of competitive advantage, as well as a currency which can be exchanged both inside the firm and with other firms. Patents are a key mechanism by which firms protect the new knowledge they produce: intellectual property rights enable a startup to enter a market, a sole inventor to create a firm in the absence of capital or customers, and a small multinational subsidiary to increase its significance in a large network of subsidiaries. This three-chapter dissertation analyzes how firms use knowledge they produce, specifically how multinational subsidiaries inventing technologies interact with their multinational headquarters and their local partners; how cutting-edge methods derived from machine learning and natural language processing can enable study of otherwise intractable problems in codifying and transferring knowledge; and how startups use patents strategically, with a focus on implications of intellectual property policy. This dissertation stands at the intersection of the fields of entrepreneurship, innovation, and machine learning.

Chapter 1 introduces a model for the relationship between the multinational firm's headquarters, its subsidiary, and the host country of the subsidiary. The model, loosely based on the gravity trade model and featuring a measure of knowledge distance introduced here, yields an answer to a longstanding topic in the multinational literature, namely whether a multinational subsidiary in a

foreign country gravitates towards its host or continues the strategy of its headquarters. The findings include a relative shift in the influence of the headquarters and host country over the subsidiary as the subsidiary grows to closer to the host, as well as the result that a highly specialized skill temporary migration visa can increase influence of the headquarters over the subsidiary when utilized. The results are relevant for both multinational managers and governments hosting multinationals.

Chapter 2 surveys key machine learning methods applied to management research, and dives especially into natural language processing applications. Applications include analyses of the patent corpus, topic modeling, and sentiment analysis. The perspectives in this chapter are relevant to the study of knowledge and broadly firm strategy, as tools from machine learning can create new measures of knowledge, transfers, and firm strategy; or improve existing ones.

The third chapter analyzes a policy shock to startup firms as a window to studying the value of reducing uncertainty in the patent examination process. Startups especially benefit from granted IP rights, as often their IP is the basis for venture funding and market entry. As the duration of the examination process is uncertain, firms treated with accelerated patenting yield significantly improved outcomes. Methodologically, the paper also adds a matching algorithm based on natural language processing to standard econometric techniques.

Table of Contents

Dedicated to the love of my life, my wife Dr. Debbie Teodorescu, and to Dr. Horia-Nicolai Teodorescu and Octavia Teodorescu, who raised me. Without their continuous support, love, and care through many years, this dissertation would not have existed.

Acknowledgments

The journey that results in a doctoral dissertation is the result of the consistent goodwill, kindness, and support of many – faculty, colleagues, mentors, friends, and family – through hundreds of interactions across several years. As doctoral students often need access to data or research sites, a dissertation is also a product of the goodwill of institutions willing to help young researchers in their pursuit of knowledge – my thanks to Harvard and the US Patent and Trademark Office. I can only hope these acknowledgments can serve as a small token of appreciation to those who were so kind to me over these years.

I would begin by thanking, from the bottom of my heart, my dissertation committee – Professors Tarun Khanna, Shane Greenstein, William Kerr, and Neil Thompson, who guided me, patiently and with great care, through this journey. During my years in the Harvard community and visits at nearby MIT, I've relied on the advice of many faculty, who graciously gave their time and guidance: Professors Juan Alcacer, Michael Toffel, Dennis Yao, Hong Luo, Andy Wu, Andrei Hagiu, Jordan Siegel, Prithwiraj Choudhury, Raffaella Sadun, Jan Rivkin, Karim Lakhani, Ryan Buell, Gary King, James Orlin, Matt Marx, Emilio Castilla, and Glenn Ellison. I'd like to thank Professor John Deighton for graciously mentoring me in both teaching and research, and for setting me on my current machine learning and big data in management path – he very generously allowed me to be a partner in teaching an MBA course, which was a foundational experience and helped direct me to my current information systems department. My special thanks to Professors Hong Luo, Andy Wu, Frank Nagle, Andrei Hagiu, Jordan Siegel, and Matt Marx, who kindly provided job market advice and time well above and beyond what a student could expect – I thank them for their generosity. I thank Faculty Assistants Maggie Kelleher, Mark Poirot, Gregory Fortier, Jillian Cunningham, and our department coordinators Sarah Owen and Sara Machunski.

Harvard Business School's doctoral programs office holds some of the most extraordinarily generous, kind, and supportive people, and I count myself lucky to have had the opportunity to know them. Whenever I needed help in securing either a research site, travel funds, or needed any advice or help, they were there for me. I wish to thank Director Jen Mucciarone, retired Director John Korn, Associate Director Marais Young, LuAnn Langan, Daria Wright, and Kathy Randel. Thank you for being so kind and supportive over the years to me and other doctoral students – we are tremendously lucky to have you. I thank Dale DeLetis and Varanya Chaubey for their patience in helping me and my graduating cohort with job market preparation, including the job market talk and flow of the job market paper, and thank the Doctoral Office for providing us with job market workshops.

Much of this work was produced thanks to generous research funding from Harvard Business School. The Harvard Business School BEST Field Research Grant enabled me to conduct my job market research at the US Patent and Trademark Office as well as start several fruitful research collaborations with researchers at the USPTO. Harvard Business School also provides a tremendous set of resources through Baker Library for doctoral students, including data and research advice from statisticians and data experts. Outside the Baker Research group, I thank Researchers Patrick Clapp and Alex Caracuzzo for data processing assistance and valuable feedback on work in progress, Dr. Xiang Ao and Andrew Marder for early discussions on methodology and valuable feedback, as well as Barbara Esty and James Zeitler.

I am so grateful to the US Patent and Trademark Office for welcoming me as a part-time colleague for almost two years of my doctorate. This was a formative learning experience for me, as well as a chance to build friendships and co-authorships which I expect will last many years to come. I wish to thank the Chief Economist Dr. Andrew Toole, previous Chief Economist Dr. Alan Marco,

Dr. Asrat Tesfayesus J.D., Dr. Nicholas Pairolero, Charles DeGrazia, Dr. Robert Kimble, Dr. Jesse Frumkin, Mr. Steve Melnick, Deputy Chief Economist Amanda Myers, Dr. Richard Miller, Sandy Phetsaenngam, Julie Chapin, Dr. Pinchus Laufer J.D., Director Jacob Betit, and the executives of the Patent Office who allow students to be a part of the USPTO through their intern/extern program.

Friends and fellow doctoral students played a vital role in my personal growth as well as in completing this dissertation. For their kindness, support, and feedback on my papers and presentations, I thank Aaron Yoon, Yo-Jud Cheng, Samantha Zyontz, Daniel Brown, Do Yoon Kim, Cheng Gao, Dr. Haris Tabakovic, Grant Donnelly, Chris Poliquin, Jasmina Chauvin, Hyunjin Kim, Stefan Dimitriadis, Young Hou, Michelle Shell, and Ohchan Kwon.

Feedback provided during research seminars, workshops, conferences, and the peer review process is essential for the development of research, and especially formative for doctoral students. I thank the participants of SMS Berlin 2016, SMS Houston 2017, INFORMS 2017, the Data Mining Workshop at INFORMS 2017, Halmstad University's CIEL innovation seminar, the HBS Strategy Doctoral Seminar, the HBS Strategy Unit Seminar, Editor Alfonso Gambardella and anonymous Strategic Management Journal referees, and the anonymous Academy of Management referees.

Oftentimes academics rely on the benevolence of private firms for access to powerful software packages. I thank Dr. Ingo Mierswa and RapidMiner for providing a free academic license for several years, as well as Dr. Anne Rozinat from Fluxicon Netherlands and Fluxicon Netherlands for waiving the licensing fees for the process mining software Disco.

Last but certainly not least, without the love and support of my family this dissertation would not have happened. I thank my wife Dr. Debbie Teodorescu and father Dr. Nicolai Teodorescu for

being so supportive during the extensive travel period necessary to develop and present this dissertation, and my grandmother Octavia Teodorescu, an attorney with many decades of wisdom and experience, who at 92 has spent many years giving me advice and taught me that honest good work is the path forward in life. My thanks as well to my extended family, who have constantly supported me despite my long periods of absence.

There are many others who have interacted me with kindness and good will in the past five years – I thank them and hope that future generations of doctoral students will both rely on such community support and pay it forward.

I would end by acknowledging that any remaining errors are my own and that any implied views in this research are mine, and do not necessarily represent the views of institutions or agencies with which I am affiliated.

Author List

Chapter 1 was co-authored with Pritwiraj Choudhury and Tarun Khanna

Chapters 2 and 3 are authored by Mike Teodorescu

Note on Figures, Tables, Appendices, and References

Labels refer to the chapter they are contained within.

Introduction

Knowledge, especially specialized knowledge that produces IP, can be used as a bargaining tool in an organization. This dissertation's first chapter, "Knowledge Flows within Multinationals – Estimating the Relative Influence of Headquarters and Host Country Using a Gravity Model," introduces a gravity model for the estimation of knowledge flows between a Multinational (MNC) subsidiary, the MNC headquarters, and the host country of the subsidiary. The MNC subsidiary that develops new products and specialized technologies has a growing relevance in the network of entities forming the MNC. Within the MNC, the subsidiary may choose to transfer knowledge to the headquarters or to develop knowledge by leveraging its local context. The subsidiary may grow closer to the knowledge-generating strategy of its headquarters, or closer to that of its host country, which may be regarded positively or negatively by the MNC. While prior studies in the international business and strategy literatures have focused on either the host-subsidiary relationship or the subsidiary-headquarters relationship, a model that allows a comparison between these two sides is needed. The study in the first chapter introduces such a model, broadly in the family of gravity models, and produces a measure of knowledge distance that can be used to compare knowledge stock at firms and subsidiaries. The focus of the study is the top twenty-five US-headquartered MNCs, by size of patenting. The findings show that as a subsidiary grows larger, the host country exerts a growing influence relative to that of the headquarters and that a certain type of immigration influences knowledge flows. A case study also shows a large MNC reacting in a manner consistent with our prediction. The methodological contributions of this chapter include a new gravity model and a set of knowledge distances applicable to MNCs, which were brought over from similarity measures used in mathematics and computer science.

Chapter 2, titled "Machine Learning Methods for Strategy Research," is primarily a review of machine learning (ML) and natural language processing (NLP) methods and their applications to management. The need for such a chapter was shown to me at one of the conferences at which the study in Chapter 1 was presented, where participants encouraged me to find applications of ML and NLP in management and attempt a comprehensive overview. This methods-focused chapter covers natural language processing methods with a focus on text analytics and machine learning methods such as classification, decision trees, boosting and cross-validation, k-nearest-neighbors, topic modeling, and sentiment analysis. The methods are presented with management examples and supplemented by references crossing multiple fields. Since the analysis in Chapter 1 introduced me to the rich data source constituted by patent text and to colleagues in the field, applications such as topic modeling of patents and corpus analysis of patent texts are included in this chapter as examples and as areas of future extension. A sentiment analysis application of corporate mottos reveals that even short bits of text provided by firms can be useful in classifying firms and potentially determining competitors beyond traditional industry code approaches. Chapter 2 also determines where the innovation field is in terms of applications of ML and NLP and helps create a foundation for some of the approaches in Chapter 3, which includes a text-based matching approach applied to startups filing for patents. In addition, the investigation reported in Chapter 2 helped to better discriminate between patents that deal with "green" technologies and other patents.

The third and final chapter of this dissertation, titled "The Need for Speed: Effects of Uncertainty Reduction in Patenting," looks at the effects of accelerating patenting on startups. The market for ideas is known as an inefficient market, primarily due to asymmetry of information and the risks of expropriation. However, the effects of uncertainty in patent grant timing on startup outcomes

have not been analyzed prior to this study. A government program by the US Patent and Trademark Office granted a "treatment" of accelerated patenting to green technologies. This provided an opportunity to determine if reduced uncertainty regarding the duration of patent pendency (the time during which an invention is under examination, a variable time that can take multiple years) would improve outcomes for startups, such as increased venture funding or earlier entry into a market, yielding higher sales and employment. This research was possible due to both access to internal US Patent and Trademark Office databases and interviews with executives and examiners in the Patent Office over an extended period. The chapter utilizes traditional econometrics methods such as difference-in-differences and coarsened exact matching and combines them with a new algorithm for constructing a control group using a classification algorithm employing concepts from natural language processing. The approach in this study may be extended to the analysis of effects of other policy changes on populations of firms or individuals for which textual data is available. Further, this study shows that a decrease of uncertainty in patenting is beneficial for startups, which may be relevant for future government programs aimed at helping small business growth.

While a dissertation is the culmination of many years of work, I recognize that it is merely the beginning of a researcher's academic output. Work is underway to extend these chapters into papers for submission in the fields of strategic management and entrepreneurship, as well as to introduce additional papers that improve upon the methods presented here.

Chapter 1:

Knowledge Flows within Multinationals – Estimating the Relative Influence of Headquarters and Host Country Using a Gravity Model

Prithwiraj Choudhury, Mike Horia Teodorescu and Tarun Khanna

ABSTRACT

To shed light on the relative influence of the headquarters (HQ) and the host country on knowledge flows to a multinational subsidiary, we use a novel methodology based on the classic gravity equation in economics and novel measures of "distance." We test our theoretical predictions using a custom dataset of patents filed by the top 25 patenting US multinationals and find that the relative influence of the HQ and the host country on knowledge flows to the subsidiary depends on the size of the subsidiary. Our findings show that as the subsidiary grows, the host country's influence on knowledge flows into the subsidiary grows faster than the influence of the HQ, which has implications for managers of MNCs. In some contexts, departure from the HQ is desirable, whereas in others an independent subsidiary may be an unwanted effect. We provide a case study of CISCO and an analysis of how certain US trade and immigration policies affect our sample. Our gravity model approach provides a new toolkit for the international business researcher, providing a means of studying the relationship between the headquarters and the subsidiary and comparing it to the relationship of the subsidiary with its host country.

KEYWORDS: MNC, Knowledge Flows, Context, Gravity Model, Cosine Similarity, Immigration

INTRODUCTION

Scholars have long hypothesized that multinational firms (MNCs) exist because of their ability to transfer and exploit knowledge more effectively and efficiently in the intrafirm context than would be possible through external market mechanisms. As Gupta and Govindrajan (2000) point out, the internalization of the intangible assets argument, originally advanced by Hymer (1960), has been widely accepted as the theory of why MNCs exist.[1] A rich empirical literature studies knowledge "inflows" and "outflows" from the perspective of MNC subsidiaries and examines how MNC headquarters (HQ) and other subsidiaries influence such knowledge flows (Feinberg & Gupta, 2004; Gupta & Govindrajan, 2000; Singh, 2008, etc.). There is also a long tradition of studying the "contexts" of local subsidiaries; scholars in this tradition have posited that MNC subsidiaries have to adapt to the local context (Birkinshaw & Hood, 1998; Ghoshal & Nohria, 1989; Kostova & Roth, 2002; Nobel & Birkinshaw, 1998; Nohria & Ghoshal, 1994; Rugman & Verbeke, 2001). Yet the seminal studies on knowledge flows within MNCs, such as Gupta and Govindrajan (2000), focus on knowledge flows between the subsidiary and the HQ or other subsidiaries and ignore knowledge flows from the host country to the subsidiary. Subsequent studies, such as Feinberg and Gupta (2004), use only US MNC subsidiary data to consider the local host country context and leave out other local firms in the host country. Given this tendency, a relatively unexplored question is how to *compare the relative influence* of the headquarters and the host country on knowledge flows into the MNC subsidiary. This gap exists even though recent literature in strategy and international business has made a strong case for studying how the local context shapes the

[1] Buckley and Casson (1976); Caves (1971), Caves, Christensen, and Diewert (1982); Ghoshal (1987); Ghoshal and Bartlett (1990); Kindleberger (1969); Porter (1986); Teece (1981).

multinational subsidiary (Dhanaraj & Khanna, 2011; Khanna, 2015; Meyer, Mudambi, & Narula, 2014; Santos & Williamson, 2015).

In this paper, we attempt to fill this gap by both theorizing and empirically studying the *relative* importance of knowledge flows to a focal MNC subsidiary from the MNC headquarters and from the host country context. An important limitation related to studying this question is the lack of an empirical technique capable of conducting an "apples to apples" comparison of how the HQ and the host country influence knowledge flows to the subsidiary. The literature on knowledge flows within MNCs has used several theoretical perspectives, including communication and transmission theory (Gupta & Govindrajan, 2000); network theory (Ghoshal & Bartlett, 1990; Hansen, 2002); cluster innovation (Alcacer & Zhao, 2012); institutional theory (Kostova & Roth, 2002); modularity (Zhao, 2006) and theories of human capital mobility (Almeida & Kogut, 1999; Choudhury, 2016; Oettl & Agrawal, 2008; Song, Almeida, & Wu, 2003, etc.). While these theories have been helpful in conceptualizing and measuring knowledge inflows to the MNC subsidiary and the influence of the MNC headquarters and other subsidiaries on such knowledge flows, they cannot produce an empirical comparison of the relative influence of the headquarters and the host country context on knowledge flows to the subsidiary. Such an empirical comparison would need to be conducted on two comparable empirical "arms." The first arm would estimate the influence of the headquarters on knowledge flows to the subsidiary, while the second would measure, using the same theoretical framework, the influence of the host country context on knowledge flows to the subsidiary.

To work around this difficulty, we use the gravity model in economics to *estimate two comparable specifications* – one that measures knowledge flows from the headquarters to a multinational subsidiary and another that measures knowledge flows from the host country context to the

subsidiary. The gravity model in economics was introduced in the trade literature with Tinbergen's (1962) work and was later formalized by Anderson (1979). Intuitively, this gravity model builds on the original model of gravity in Newtonian mechanics, where the force of attraction between two bodies is proportional to the masses of the two bodies and inversely proportional to the square of the distance between those bodies. The equivalent form of the gravity equation in the trade literature substitutes the force of attraction as a dependent variable with a measure of trade, such as the dollar flow of traded goods between two countries (Anderson, 1979) in relation to the masses of the two countries as measured in GDP (Mátyás, 1997), and in relation to a distance variable, often specified as the geographic distance between the two trading regions (Anderson, 1979).[2] The gravity model is highly generalizable, as long as one has sensible measures of flow (force), mass, and distance. Consequently, it is now being used in various fields beyond trade economics.[3] For example, Lewer and Van den Berg (2008) developed a gravity model of immigration. In this case, the flow is immigration, the mass variables are the populations of the pairs of countries, and the distance is defined traditionally as geographic distance.

We apply a generalized form of the gravity model to study the relative influence of the MNC headquarters and the local context of the host country on knowledge flows to the MNC subsidiary. To conduct this empirical analysis, we estimate two comparable equations (which serve as the *two comparable empirical arms* of our analyses). The first equation uses the "mass" of the

[2] Bergstrand (1985) further expanded the theoretical foundations of the gravity model in the trade literature by deriving it from a general equilibrium model. While gravity equation-based regression models have generally been estimated via OLS (e.g., Mátyás, 1997), the OLS estimator has been shown to be biased in the case of heteroscedasticity (Silva & Tenreyro, 2006). Silva and Tenreyro (2006) instead proposed a Poisson pseudo-maximum-likelihood estimator for the cases where the errors are heteroskedastic. Additional econometric developments have been made, for example by Baier and Bergstrand (2008).

[3] Other relevant papers that use the gravity model include Anderson and Wincoop (2003), who focus on the effect of national borders as trade barriers and determine that borders substantially reduce trade by 20-50%. Waugh (2010) focuses on determining trade flow asymmetries between countries based on differences in the standards of living of the trading countries. Summary (1989) used measures of political factors between trading partners as additional independent variables to augment a regression model based on the gravity equation.

headquarters, the "mass" of the subsidiary and the "distance" between them; the second equation uses the "mass" of the host country, the "mass" of the subsidiary and the "distance" between them.

To estimate these two equations and compare the marginal effects, we need relevant measures of knowledge flows from the MNC headquarters and the host country to the MNC subsidiary, measures for the "mass" of the knowledge stock at the MNC headquarters/in the local context and measures for the "knowledge distance" between the focal subsidiary and the MNC headquarters/local context. We quantify knowledge flows through the use of patent citations, a widely accepted measure in the innovation literature (Jaffe *et al*., 1993). To estimate the "mass" variables, we use the stock of patents filed by an entity. Measuring "distance" between MNC entities (HQ, subsidiary and host country context) presents unique challenges: the gravity model in economics has typically used geographic distance to model trade flows. However, the literature in knowledge flows has shown that social and ethnic ties between inventors influence knowledge flows between locations (Agrawal *et al*., 2006; Agrawal *et al*., 2008). Given this, geographic distance may be less salient in determining knowledge flow between MNC entities. Instead, we use several novel measures of distance based on cosine similarity measures. As we explain in detail later, we also employ a novel measure of distance based on the Bhattacharya coefficient, Hellinger affinity, and "fidelity similarity" (Deza & Deza, 2015). Using these measures, we separately measure the effect of the MNC headquarters and the local context on knowledge flows to the subsidiary.

To conduct this analysis, we created and used a novel dataset of US patents filed by the top 25 US headquartered multinationals at the US Patent and Trademark Office (USPTO) over a period of 7 years (2005–2011) to calculate measures of knowledge distance between entities within MNCs. The dataset included an assignee disambiguation task and was not readily available, given that

over 60 countries are included for the top-inventing US-headquartered MNCs. These patents were filed at either the headquarters or any of the subsidiaries of these firms around the world. We coded the location from which each patent was filed. We also created a "mass" measure for each headquarters, host country and subsidiary from 2005 to 2011 and created distance measures using custom-written software code. We report several results. First, we validate the gravity model specifications separately for the influence of the headquarters on knowledge flows into the subsidiary and the influence of the host country on knowledge flows into the subsidiary. These results are robust to the inclusion of several controls. Second, we conduct marginal analyses to establish that the relative influence of the HQ and the host country on subsidiary knowledge flows depends on the size of the subsidiary. As the size of the subsidiary grows, the host country's influence on knowledge flows into the subsidiary grows faster than the influence of the headquarters. This indicates possible heterogeneity across MNC subsidiaries in the relative importance of the host country context and the headquarters. Our results indicate that the host country context is more important for subsidiaries that have a greater stock of patented innovations. Finally, we study the mechanisms behind our findings and present evidence that immigration policy positively affects knowledge flows.

The chapter is organized as follows: in section 2 we overview the fundamentals of MNC theory that we build upon and introduce a gravitational model for knowledge flows and the hypotheses; in section 3 we summarize our data and empirical approach; the results are presented in section 4; in section 5 we provide a discussion of the data and findings, and section 6 summarizes our conclusions.

THEORETICAL MODEL AND HYPOTHESES

Theoretical Foundations – The Headquarters

The earliest models of the MNC view the organizational structure as a "centralized hub" (Bartlett, 1986), where the HQ directs resources, tasks, and relationships to the MNC subsidiaries. Further, Ghoshal (1986) recognized that subsidiaries have an advantage in specialization over the HQ, in that subsidiaries have closer contact with their host country than the HQ might have. For instance, works such as Caves (1971), Doz, Bartlett, and Prahalad (1981), and Hymer (1960; 1976) have recognized that managers native to the country where an MNC subsidiary is located possess knowledge and relationships external to the HQ that give them an advantage over the HQ. Knowledge transfers are quintessential to the existence of the MNC and serve as currency for the organization. The "interorganizational network" view of the MNC is introduced in Ghoshal and Bartlett (1990) – the organizational units of the MNC, which include its subsidiaries and its HQ, are embedded in an "external network" that consists of all the entities the MNC interacts with. Within this interorganizational network, the HQ may assign different strategic roles to its MNC subsidiaries. There is a rich literature stream establishing the key role of the HQ in the MNC, including Andersson, Forsgren, and Holm (2002), Björkman, Barner-Rasmussen, and Li (2004), Dacin, Beal, and Ventresca (1999), Ghoshal and Bartlett (1990), Ghoshal and Nohria (1989), Gupta and Govindarajan (1991), and Nell and Ambos (2013). Ghoshal and Bartlett (1990) see the MNC as "somewhere between […] unitary and federative structures" (Ghoshal and Bartlett 1990, p. 607), meaning that in some MNCs, the goals are set and the decisions are made with full authority by the HQ, with the subsidiaries following, while in other MNCs, the subsidiaries are given the choice of whether to ratify the decisions handed to them from the HQ. In the same line, Ghoshal and Nohria (1989) argue that in the HQ-subsidiaries relationship, context plays a key role,

determining the degree of the centralization of authority, from a dissolution of centralization up to structures that are similar to clans and integrative structures. In the network view of Ghoshal and Bartlett (1990), the HQ creates value for the network and for the subsidiary through its facilitation of resource transfers within the organization and its ability to assign and oversee strategic roles for the various subsidiaries. Andersson *et al*. (2002) and Dacin *et al*. (1999) found that the HQ maintains an external network of ties to outside actors, even when such ties are duplicates of one of its subsidiaries' external links. The work in Nell and Ambos (2013) shows that the HQ can in fact generate benefits for the MNC subsidiaries even if such external ties are shared across entities within the MNC, and that the benefits provided by the HQ through its external ties are stronger for younger subsidiaries and stronger for the subsidiaries when the HQ is more embedded in its subsidiary's local context. Gupta and Govindarajan (1991) introduced knowledge as the key differentiator between subsidiaries and the source of their organizational power. A subsidiary may either be a "receiver" of knowledge or a "provider" in any given knowledge transaction, and those subsidiaries that primarily generate knowledge command more authority in the organization.

In any exchange process, including a knowledge flow, the flow is influenced by the exchange partners. An ample corpus of literature has been devoted to the influence of knowledge flows in MNCs and on the influences of the HQ and their subsidiaries on the inbound and outbound flows. This broad variety of links between HQ and subsidiaries should necessarily be reflected in the knowledge flows from HQ to subsidiaries, from imposed flows to local interest-based flows. A detailed analysis in Gupta and Govindarajan (2000) shows that "knowledge inflows into a subsidiary is positively associated with richness of transmission channels and motivational disposition to acquire knowledge" (Gupta and Govindajaran, 2000, p. 473); these are both influenced by the HQ, which may invest in the increase of the bandwidth of the transmission

channels and force goals for the subsidiary that would motivate it to absorb knowledge. However, as Narula (2014) shows, maintaining a wide bandwidth (high capacity channels) for knowledge transfers with the subsidiaries may prove costly. As Gupta and Govindarajan (1991) and Narula (2014) show, different subsidiaries may have very different types of inward knowledge flows and different levels of control exerted by the MNC, which, as argued first in Gupta and Govindarajan (1991), may be due to the different contexts where the subsidiaries operate. Among the control means the HQ may use to increase knowledge transfer to itself, Björkman *et al*. (2004) identify in the first place "the specification of knowledge transfer as a criterion of subsidiary performance" (p. 446), which may push the subsidiary to transfer more knowledge to the HQ and other subsidiaries – and to receive more knowledge from other subsidiaries. Björkman *et al*. (2004) also found "strong positive relationship between subsidiary stock of knowledge and knowledge transfer" (p. 452) inside the MNC, especially to other subsidiaries. Björkman *et al*. (2004), Gupta and Govindarajan (2000), and O'Donnell (2000) identify managerial socialization as an opportunity to bring the subsidiary closer to the HQ vision, including through "international training programmes, by establishing international task forces and committees, and by encouraging visits across MNC units" (Björkman *et al*., 2004, p. 451). In sum, the fact that the HQ influences knowledge flows and decisions at the level of the subsidiary is a well-researched topic.

Theoretical Foundations – The Subsidiary

It has been shown that subsidiaries with higher knowledge output and more connections to their local context are more valuable to the MNC and its HQ (Almeida & Phene, 2004). Gupta and Govindarajan (1991), building on the TCE literature, define this type of subsidiary as playing a "Global Innovator role," in which the subsidiary's benefit to the MNC is driven by its unique knowledge-generating potential, knowledge that is used as currency in exchanges with other units

within the MNC organization, as well as by the lower intra-organizational knowledge transfer transaction costs. However, Björkman *et al*. (2004) argue that "it may […] be in the subsidiary's self-interest not to transfer knowledge" (p. 444), as this very knowledge may be the *raison d'être* of the subsidiary and its source of competitive advantage within the organization. Subsidiaries do make decisions that do not maximize "corporate performance" (Björkman *et al*., 2004), such as withholding knowledge, if sharing that knowledge risks the standing of the subsidiary within the larger MNC. Monteiro, Arvidsson, and Birkinshaw (2008) have argued that "some subsidiaries are isolated from knowledge transfer activities within the multinational" (p. 90), because they do not belong to the units "perceived to be highly capable" (p. 90), or because of the low "levels of communication and reciprocity" (p. 94). These "levels of communication" are directly connected to the high-capacity channels described by Narula (2014) and to the richness of transmission channels emphasized by Gupta and Govindarajan (2000). Studying the product flow only, Birkinshaw and Morrison (1995) found in a study focused on configurations of the MNCs that the parent-subsidiary relationship differs substantially for "world mandate subsidiaries" and local subsidiaries, with the former experiencing a significantly larger strategic autonomy, which may positively influence the ability of the subsidiary to choose its level of knowledge absorption and knowledge sources (Birkinshaw & Morrison, 1995, p. 744). However, this finding is not fully in line with the findings of the other authors previously cited. Birkinshaw and Hood (1998) draw attention to the fact that there is an "enormous variety of subsidiaries in existence" (p. 773), and take two viewpoints, that of network theory and that of the decision process in large organizations, concluding that the subsidiaries are continuously evolving as elements of a network, sometimes going beyond a strict dyadic relationship with the HQ, with the evolution being propelled by the "underlying capabilities" of the subsidiaries (Birkinshaw & Hood, 1998, p. 782) and by the degree

of autonomy they have to make decisions and take initiative as entities, not just as a part of the complex organism that is the MNC. The evolution of the international subsidiaries may gravitate to independence, and consequently to a declining role in the network of the MNC, as emphasized in Birkinshaw and Hood (1998). This tension, between the mandate given to the subsidiary by its parent and the subsidiary's evolving charter, is fueled by the subsidiary's growing capabilities in its local network. Subsidiaries are no longer "resource seeking, market seeking, or efficiency seeking" entities (Birkinshaw & Hood, 1998, p. 773); instead, they create their own dynamic capabilities (Nelson & Winter, 1982) and become sources of competitive advantage for their parent organizations. In the recent MNC literature, the role of the subsidiary is a fluid one, and the subsidiary can dynamically change its scope and importance in the organization. This aspect of the subsidiary is modeled with the gravity approach we propose here.

Theoretical Foundations – The Host Country

One of the earliest views of the role of host country in the literature is that of the Product Life Cycle model (Vernon, 1966), in which locating a manufacturing plant abroad is a natural part of the product life cycle: later-stage technologies can benefit from the lower cost of manufacturing in a location other than where they were invented. The move abroad in this model is driven by production costs, and the subsidiary is always under the directive of the HQ. The later network model of Ghoshal and Bartlett (1990) and Rugman and Verbeke (1992) allows for ties between the subsidiary and the host country that are inherently valuable, external to the HQ and costly for the HQ to develop abroad. The most recent studies of knowledge flows and innovation in MNCs build upon this growing role and influence of the host country for the subsidiary, and introduce the perspective that subsidiaries evolve and might turn to their host countries, as opposed to the HQ, to define their charter and role. For instance, Mudambi, Pedersen, and Andersson (2014) find

that a subsidiary may turn toward to the national agents of innovation as a replacement of the HQ as a source of ideas and platform for ideas exchange. Mudambi *et al*. (2014) stress that "there is evidence that headquarters' fiat power in MNCs is not absolute" (p. 102) and that the MNC authority may never be fully exerted or may degrade during time, a finding applicable to the power of innovation and to the control of knowledge flows. Technological knowledge assets in the host country may become a basis of the subsidiaries' power. In an analysis of the influence of the MNC and the host country on innovation in the process of knowledge creation by subsidiaries, Almeida and Phene (2004) argue that foreign subsidiaries of the MNC evolve in two contexts: the context of the MNC network and that of the local context. Accordingly, several factors play a role in the generation of innovation by subsidiaries. Two of these factors are identified by Almeida and Phene (2004) at the local level, namely local technological richness and the strength of the links the subsidiary has with local entities. O'Donnell (2000) finds that a subsidiary with "a high level of specialized information [the] headquarters does not have" gains a "strategic role" in the MNC (p. 527); such a role provides favorable terms to the management of the subsidiary and is inherently valuable. In summary, subsidiaries have the option to turn toward the local context when the host country provides more resources and more favorable incentives for innovation and knowledge exchange than does the HQ; this behavior is well observed. The vast literature on other reasons host countries matter to the evolution of subsidiaries cites host country governmental pressure (Doz *et al*., 1981), local customers' influence (Doz *et al*., 1981), the national institutional context and local educational system (Almeida & Phene, 2004; Caves, 1974), and the level of development of the domain of the subsidiary in the host country (Singh, 2007). Therefore, the power of the HQ over the subsidiary is not absolute and can be negotiated by a subsidiary through its development of knowledge and technological assets and by the strengthening of its ties with the host country.

Despite the rich literature on MNCs, which studies both the relationship between the HQ and its subsidiaries and that between the subsidiary and its host country, we lack an empirical tool to compare the knowledge flows between these entities. Specifically, a research problem that remains unsolved is the comparison of the relative influence of the HQ and the host country on knowledge flows to subsidiaries. We also need to study the factors that might shape the relative influence. A new approach is needed to compare back-to-back the influences of the HQ and the host country on knowledge spillovers to subsidiaries.

This study makes several contributions in two directions. First, we propose the gravitational model as a tool to compare the influences of the HQ and the host country on innovation in subsidiaries. Second, we perform the study using a new database including data collected for a large number of US-based MNCs with subsidiaries in a large number of countries.

The Gravity Model

As discussed earlier, we propose a gravity model to analyze the knowledge flows within an MNC and to *estimate two comparable specifications* – one measuring knowledge flows from the headquarters to a multinational subsidiary and the second measuring knowledge flows from the host country context to the subsidiary. The traditional gravity model in the trade literature establishes an inverse proportionality between a trade variable (such as trade flow between two countries) and the physical distance between the countries, as well as a direct proportionality of the trade variable to two "mass" variables that represent measures of the trading capacity of the two countries, such as their respective GDPs. An example is (Mátyás, 1997, p. 363):

$$\ln(EXP_{ijt}) = \alpha_i + \beta_1 \cdot GDP_{it} + \beta_2 \cdot GDP_{jt} + \beta_3 \cdot d_{ijt} + \epsilon_{ijt}$$

The choice of a distance function and the economic measure equivalent to mass has varied depending on the application. In Mátyás (1997), for example, the dependent variable was the volume of trade between the two countries and the two masses were the populations of the two countries. In Bergstrand (1985), the distance was the physical distance between economic centers and the masses were the GDP values in year *t* of the two countries. Numerous alterations to the model are possible, including the addition of other independent variables such as foreign currency reserves (Mátyás, 1997), without changing the key components of the gravity equation, specifically the two measures of mass and the measure of distance. This flexibility and the derivation of the regression model encourage wide applications, specifically enabling various definitions of distance (there are infinite distance functions beyond the popularly used geographic distance) and various options for the independent variable. To shed further light on these variations, one may look at the Newtonian physics formula that lies at the origin of the gravity-based regression model.

This regression model is analogous to the gravitational attraction force in Newtonian mechanics transformed into a linear form appropriate for a regression through a logarithm. A gravitational force between two bodies *i* and *j* is defined as:

$$F_{\{i,j\}} = \gamma \cdot \frac{M_i \cdot M_j}{d_{ij}^2}$$

Taking the log results in the traditional gravity regression model in the economics trade literature, where the "force" is replaced with a measure of economic flow between the two entities *i* and *j*:

$$\ln F_{ij} = \gamma + \beta_1 \cdot M_i + \beta_2 \cdot M_j + \beta_3 \cdot d_{ij} + \epsilon_{ij}$$

We look at research and development subsidiaries, and our definition of subsidiary here is that of a company branch located in a different country from that of the headquarters that produces patents filed at the USPTO. As our entire sample consists of US-headquartered companies, all subsidiaries are located outside the US. We hypothesize that an innovation produced by an MNC subsidiary is influenced by two gravitational-like forces: a force generated by the MNC headquarters and acting upon the subsidiary and a competing force generated by the subsidiary's host country and acting upon the subsidiary. Using a celestial mechanics metaphor, a subsidiary could be said to gravitate around either its headquarters or its host country. Figure 1 shows an intuitive schematic of these competing gravitational forces.

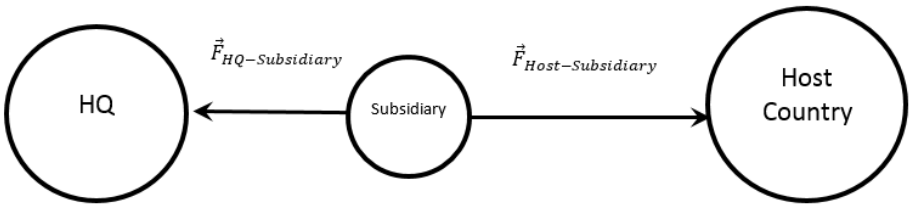

Figure 1. Schematic of two competing gravitational forces on the subsidiary.

Modeling the Subsidiary-Headquarters Relationship

The trading partners in this case are the subsidiary and the headquarters. Following the gravity model, we define two masses: the mass of the headquarters of the firm (m_{HQ}) and the mass of a subsidiary of the firm (m_S). The mass is defined as the number of issued patents originating from that entity's location within a year. The measure of trade in our case is the count of patent citations. While traditional gravity models define the distance between the trading partners as geographical distance, such a measure does not necessarily apply to knowledge flows, and thus we introduce a new measurement of knowledge distance based on patent citations, as described in the methods

section. In this knowledge flow measure, a flow from an entity to the subsidiary is evidenced through the subsidiary patents' citations of the patents of that entity. For every pair (headquarters, subsidiary) and every year, we thus calculate all citations of the patents of the headquarters in the patents of the subsidiary and define the result as our dependent variable. We hypothesize that the knowledge flow relationship between an MNC headquarters and a foreign-located subsidiary follows a gravitational model. Specifically, we make the following baseline hypothesis:

Hypothesis 0a: Knowledge flows from the headquarters to the subsidiary increase in direct proportion to the stock of knowledge at the headquarters.

Knowledge flows from the headquarters to the subsidiary increase in direct proportion to the stock of knowledge at the subsidiary.

Knowledge flows from the headquarters to the subsidiary decrease proportionally to the square of the knowledge distance between the headquarters and the subsidiary.

Modeling the Subsidiary-Host Country Relationship

The trading partners in this case are the subsidiary and its host country. We define two masses: the mass of the firm's host country (m_C) and the mass of the firm's subsidiary (m_S). The measure of trade is again citation-based. Specifically, for every pair (host country, subsidiary) and every year, we calculate all citations of host country-originated patents in the subsidiary-originated patents and define the result as our dependent variable. We use the same distance measure models as in the previous case. We thus posit that the relationship between a subsidiary and its host country follows a gravitational model, yielding the following baseline hypothesis:

Hypothesis 0b: Knowledge flows from the host country to the subsidiary increase in direct proportion to the host country's stock of knowledge.

Knowledge flows from the host country to the subsidiary increase in direct proportion to the subsidiary's stock of knowledge.

Knowledge flows from the host country to the subsidiary decrease proportionally to the square of the knowledge distance between the host country and the subsidiary.

Relative Effects and Moderators – Absorptive Capacity and HQ Country Immigration

Cohen and Levinthal (1990) introduced the concept of "absorptive capacity" as "the ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends" (p. 128). Absorptive capacity is inherently about innovation and the ability to innovate, and exists both at the individual and organizational levels. Cohen and Levinthal (1990) identify "knowledge diversity" (p. 131) and "individuals who stand at the interface of either the firm and the external environment or at the interface between subunits within the firm" (p. 132) as sources of absorptive capacity. Knowledge diversity can be accomplished by the MNC by launching a subsidiary in a location with either individuals or a knowledge set different from those at its HQ, as the MNC can incorporate external knowledge and cumulatively innovate upon it, as well as create connections between internal and external knowledge to innovate. Cohen and Levinthal (1990) identify such "novel linkages and associations" (p. 133) as achievable at both the interpersonal and organizational levels. This is applicable to our analysis; for instance, the MNC may create interaction channels between its HQ and its subsidiaries. Recent MNC literature has departed from the "liability of foreignness" view and has been finding that firms exploit differences between countries to "develop unique and potentially valuable capabilities, and foster learning and innovation" (Stahl, Tung, Kostova, & Zellmer-Bruhn, 2016, p. 623). Indeed, the sign of an organization with high absorptive capacity is its ability to pursue "emerging technological opportunities" (Cohen & Levinthal, 1990, p. 137), as opening subsidiaries in countries with high-

skill workers and high-technology industries would be called. In an example we will describe in a later section on marginal effects, CISCO opened a second R&D HQ in India, essentially an HQ for Asia, increasing its absorptive capacity and its responsiveness to emerging technologies. Zahra and George (2002) built upon the concept of absorptive capacity through the lens of dynamic capabilities to define four complementary capabilities that comprise absorptive capacity; of these, acquisition (finding and acquiring external knowledge that is valuable to the firm) and assimilation ("routines and processes that allow [the firm] to analyze, process, interpret and understand the information obtained from external sources," Zahra and George 2002 p. 189) are the means by which the MNC acquires knowledge from its subsidiaries. A subsidiary rich in knowledge external to MNC's HQ innovation portfolio and channels of managerial collaboration between the subsidiary and its HQ are both sources for increasing absorptive capacity.

As a subsidiary grows in size, it develops knowledge and relationships external to the MNC that grant it a degree of independence from the HQ. The foreign subsidiary develops those processes necessary to leverage knowledge specific to its local context and external to the MNC HQ, a capability that leads to stronger knowledge flows from the host country as the absorptive capability of the subsidiary grows. In our framework, this leads to the following hypothesis:

Hypothesis 1: As the subsidiary's knowledge stock increases, the rate of growth of its absorptive capability for knowledge from its local context grows faster than its capability to absorb knowledge from the headquarters.

Whether the subsidiary turns away from the MNC HQ in terms of sourcing its knowledge may have implications for both international business researchers and managers, because while some MNCs, such as CISCO, prefer a subsidiary with very high degree of autonomy, going as far as to establish a second R&D HQ (CISCO Bangalore), others prefer a tighter integration between HQ

and subsidiary R&D. For the latter, the literature predicts that tighter integration with the HQ can be achieved through programs for the subsidiary managers. As theory suggests, managerial socialization, especially through the travel of subsidiary managers to other locations (O'Donnell, 2000; Björkman *et al*., 2004) and high-level managerial training, are means by which the HQ can bring the subsidiary closer to its interests. This reasoning, along with the unique structure of the US immigration system (as all our corporate HQs are in the US), led us to look at temporary migration of only high-level managers and experts as a potential measurement of managerial socialization and as a mediator of HQ-subsidiary knowledge flows:

Hypothesis 2: The strength of the knowledge flows from the subsidiary to the headquarters is mediated by the temporary migration of high-skill managers and specialists of extraordinary ability from subsidiary countries.

DATA AND METHODS

Sample

Our target data consist of a list of the patents issued to the top 25 US-headquartered patentees as measured by the volume of patents issued at the USPTO. We created a unique dataset of the USPTO patents filed by all of the subsidiaries of these MNCs from 2005 to 2011 (inclusive). The information collected comprised all patent bibliographic information, patent citation data, and patent textual data. Because the readily available existing patent datasets did not disambiguate the locations and assignees or the patent textual data for recent years, we created a custom dataset. Our raw dataset comprises all patents issued by the USPTO between the dates of January 1st, 2005 and December 31st, 2011, yielding 1.27 million patents and 69.3 million citations. We extracted this data from Thompson Innovation and enhanced it with string processing techniques to

disambiguate company names and inventor locations. The USPTO defines the origin of a patent as the location of residence of the first inventor.[4] Following the same definition, we took additional processing steps to improve the accuracy of the inventor location data. Our dataset was stored in a SQL database and further processed with a custom built 2000-line C# program to construct all of the variables in our proposed gravity model.

Variables

For the relationship between subsidiary and headquarters, the dependent variable is citations from subsidiary to headquarters within the same company; the independent variables are a measure of the patent output of the subsidiary per year (mass of subsidiary), a measure of the patent output of the headquarters (mass of headquarters), and a measure of the knowledge distance between the innovation outputs of the subsidiary and the headquarters. For the relationship between subsidiary and host country, the dependent variable is the subsidiary's citations of host country patents, the mass of the headquarters is replaced with the mass of the country, and the knowledge distance is the distance between the innovation outputs of the subsidiary and the outputs of its host country.

Specifications

Subsidiary-headquarters relationship. Our model is a gravity-like model where the dependent variable is the count of a subsidiary's citations of patents filed by the company's headquarters, and our independent variables are, as in a standard gravity model, mass of subsidiary, mass of headquarters, and distance between subsidiary and headquarters. The regression for the case of the subsidiary-headquarters relationship using distances is as follows:

[4]U.S. PATENT AND TRADEMARK OFFICE Patent Technology Monitoring Team (PTMT), "Patenting By Geographic Region (State and Country) Breakout By Organization." Accessed 1/4/16. http://www.uspto.gov/web/offices/ac/ido/oeip/taf/stcasg/inx_stcorg.htm

$$BK_{CITES_{ijt}} = \beta_0 + \beta_1 \cdot m_{ijt} + \beta_2 \cdot m_{i_{HQ}t} + \beta_3 \cdot d_{ijt} + \beta_4 \cdot YEAR_t + \beta_5 \cdot FIRM_i \ (1),$$

where m_{ijt} represents the log of the mass of the subsidiary *j* of company *i* in year *t*, $m_{i_{HQ}t}$ represents the log of the mass of the headquarters of company *i* in year *t*, and d_{ijt} represents the log of the knowledge distance between subsidiary *j* and the headquarters of company *i* in year *t*, and YEAR and FIRM are fixed effects dummies (all mass variables are stocks of patents). Another equation is used in which similarity measures replace the distance variable above. Due to the limitations of our data, we are unable to distinguish different subsidiaries of the same firm in the same country and therefore consider all R&D activity originating from a firm in a given country to be coming from one subsidiary.

While traditional gravity models define the distance between the trading partners as geographical distance, such a measure does not necessarily apply to knowledge flows. In a typical trade model, geographic distance matters (because of transport costs, border crossing costs, etc.), and in consequence the literature uses the physical distance between the trading partners. Although any citation involves some search cost, the advent of online search engines means that the cost does not depend on geographical distance, discounting the web lag time in accessing the main patent aggregators, such as the European Patent Office or the USPTO, from various international locations. We therefore propose a non-geographic measure of distance that is based on knowledge similarity between the trading parties. Specifically, our gravity model is based on knowledge distance, as in (1), and includes an additional specification based on knowledge similarity measures, shown in (2) below. Both distances and similarity measures are popular in data mining. Model (2), based on a similarity measure σ_{ijt} between subsidiary and headquarters, is:

$$BK_{CITES_{ijt}} = \beta_0 + \beta_1 \cdot m_{ijt} + \beta_2 \cdot m_{i_{HQ}t} + \beta_3 \cdot \sigma_{ijt} + \beta_4 \cdot YEAR_t + \beta_5 \cdot FIRM_i (2)$$

Regarding the dependent variable choice, a traditional gravity model applied to trade would not involve a discrete trade variable. In our case, the back-citation count (our knowledge flow, or trade) is a discrete variable, and therefore a count regression model is more appropriate. We are preserving all other essential aspects of a gravity model (all of the mass and distance variables that remain are logged). In terms of notation, all lower-case variables have already been transformed through a logarithm.

Subsidiary-host country relationship. A gravity model specification for the subsidiary-host country relationship is:

$$BK_{CITES_{ijt}} = \beta_0 + \beta_1 \cdot m_{ijt} + \beta_2 \cdot m_{C_{ijt}} + \beta_3 \cdot d_{ijt} + \beta_4 \cdot YEAR_t + \beta_5 \cdot FIRM_i \ (3),$$

where m_{ijt} represents the log of the mass of the subsidiary *j* of company *i* in year *t*, $m_{C_{ijt}}$ represents the log of the mass of the host country[5] of subsidiary *j* of company *i* in year *t*, d_{ijt} represents the log of the distance between subsidiary *j* of company *i* and its host country in year *t*, and YEAR represents dummies for year fixed effects. The dependent variable is again measured as the subsidiary patents' citations of patents in the host country (not patented by the same firm as the subsidiary).

The model based on a similarity measure[6] σ_{ijt} between a subsidiary and its host country is:

$$BK_{CITES_{ijt}} = \beta_0 + \beta_1 \cdot m_{ijt} + \beta_2 \cdot m_{i_{HQ}t} + \beta_3 \cdot \sigma_{ijt} + \beta_4 \cdot YEAR_t + \beta_5 \cdot FIRM_i \ (4)$$

[5] We note that in some cases, the mass of the country as defined in terms of stock of patents is comparable to the mass of the headquarters of some of the firms.
[6] A similarity measure is roughly the equivalent of the inverse of a distance. The mathematical details are explained in section 3.4.

While we have already defined the masses and the dependent variables for the two relationships of interest (subsidiary-headquarters and subsidiary-host country), we still must define a measure of knowledge distance to properly specify the gravity model.

Measures of Knowledge Distance

Cosine similarity. We implemented several measures of knowledge distance, starting with cosine similarity. The cosine similarity measure is defined as the cosine between two identically-sized vectors. Given vectors $\vec{u}, \vec{v}$, the cosine similarity measure σ is obtained from the dot product of the two vectors:

$$\sigma = \cos(\vec{u}, \vec{v}) = \frac{\sum_{i=1}^{N} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \cdot \sqrt{\sum_{i=1}^{N} v_i^2}}$$

where $\sigma \in [-1,1]$, $\sigma = -1$ for completely opposite vectors (angle of 180^{O}), $\sigma = 0$ for orthogonal vectors, and $\sigma = 1$ for identical vectors. The cosine similarity measure is used in the management innovation literature (Kay, Newman, Youtie, Porter, & Rafols, 2014) both for measuring the similarity of patenting activity through patent class counts and for textual analysis. The cosine similarity is widely used in the fields of mathematics and computer science and is one of the most popular similarity measures. Cosine similarity is the measure used in our baseline model.

We use the cosine similarity as follows: for every year, within each firm we create a vector of patent class counts for every subsidiary and headquarters representing all USPTO patent classes for utility patents (slightly over 400 classes). We weigh the patent class counts based on the total number of patents issued to the subsidiary in that year. (Comparing the raw patent counts per patent class between the subsidiary and the headquarters is not appropriate without taking into account

the different total patent outputs of the two entities.) The k^{th} element of the vector of the weighted patent class counts for company *i,* subsidiary *j*, year *t*, and patent class *k* is computed as follows:

$$v_{ijtk} = \frac{Count(issued\ patents\ company\ i\ subsidiary\ j\ year\ t\ in\ class\ k)}{\sum_{k=1}^{M} Count(issued\ patents\ company\ i\ subsidiary\ j\ year\ t\ in\ class\ k)}$$

The patent classes are not sequential. They range from a class of 2 to a class of 987, but with gaps, the total number of patent classes is slightly over 400, so *M* in the above is about 400. In the case of our model, patent counts are never negative numbers, and so the interval for our similarity measure is $\sigma \in [0,1]$.

This measure is suitable for use along with the previously defined distances because subsidiaries generally follow the research agenda of the headquarters. Specifically, the subdomains of R&D found in the subsidiary's patents are typically a subset of the subdomains of the headquarters, as measured in patents issued per class of the subsidiary and headquarters. We ran regressions (2) and (4) with cosine similarity and reported the results as the baseline model, labeled **model 1.** As this is a similarity measure, the expected coefficient should be positive.

Similarity measure to distance function. A distance function can be intuitively thought of as the inverse of a similarity measure, specifically any transformation *d* satisfying $d(x,x) = 0, d(x,y) = d(y,x)$, $d(x,y) = 0 \Leftrightarrow x = y$, and $d(x,z) \leq d(x,y) + d(y,z)$. We used three standard transformations from a similarity measure to a distance function for additional models (labeled as models 2–4 both below and in the results tables):

$$d_1 = -\log(\sigma)\ \textbf{(2);}\ d_2 = \log(1000 - 1000 \cdot \sigma)\ \textbf{(3);}\ d_3 = \log\left(1000 - 1000 \cdot \sqrt{\sigma}\right)\ \textbf{(4),}$$

where d_1 is simply a standard irrespective of the minimum value of σ, and d_2 and d_3 represent two additional transformations from a similarity measure to a distance function, taking into account

that the lowest non-zero value of similarity in our dataset is of the order of 0.001.[7] These three distances are isomorphic with the cosine similarity in (1), do not change the significance of estimates, and are constructed to place our model into a standard gravity equation, which generally uses a measure of distance between entities.

Bhattacharya coefficient – fidelity similarity. We also used the similarity measure known as the Bhattacharya coefficient, Hellinger affinity, or "fidelity similarity" (Deza & Deza, 2015). For two vectors, this is defined as:

$$\rho(\vec{u}, \vec{v}) = \sum_{i=1}^{N} \sqrt{u_i} \cdot \sqrt{v_i}$$

The fidelity similarity works better than the cosine similarity for vectors with components that are close together; it results in a more compact interval. The fidelity similarity is depicted as **model 5** in both the subsidiary to headquarters regression models and the subsidiary to host country regression models.

[7] In our balanced panel dataset covering 25 companies over years 2005-2011, we include all countries where patenting activity occurs. Roughly 50% of the data points contain a zero-patenting subsidiary (thus mass of subsidiary is 0), which implies that the patent counts vector for those subsidiaries is null, resulting in a similarity value of 0.

RESULTS

Summary Statistics

Table 1 represents the summary statistics for the variables in the subsidiary to headquarters regressions (five models, with models 1 and 5 based on similarity measures, and models 2-4 based on distance measures). Table 2 represents the summary statistics for the variables in the subsidiary to host country regressions, following the same five models as in the headquarters to subsidiary case, and sharing one variable with Table 1 (logged mass of subsidiary). Notice that in Tables 1 and 2, the mass measures are plausible (logarithmic scale): the subsidiaries are orders of magnitude smaller than the headquarters, whereas the largest countries are a few orders of magnitude larger than the headquarters (with the country maximum being about 100 times larger than the company maximum). The similarity between the vectors of patenting of headquarters and subsidiaries is typically larger than the similarity between subsidiaries and host countries, which is to be expected.

Table 1: Summary Statistics for Variables in Subsidiary to HQ Regressions

	(mean)	(sd)	(min)	(max)
Cites Subsidiary to HQ	.9570815	4.547072	0	82
Log(Subsidiary Mass)	.8209908	1.10791	0	5.204007
Log(HQ Mass)	6.414829	1.033538	1.791759	8.483843
Cosine Similarity	.1780619	.2593527	0	1
Distance d_1	4.275926	8.166304	.0145975	6.907755
Distance d_2	6.61518	.5400273	2.673558	6.906755
Distance d_3	6.391382	.7381062	1.984054	6.875622
Fidelity Similarity	.1726259	.2269262	.001	.8693181

Table 2: Summary Statistics for Variables in Subsidiary to Country Data

	(mean)	(sd)	(min)	(max)
Cites Subsidiary to Country	.4325567	2.748737	0	59
Log(Subsidiary Mass)	.8209908	1.10791	0	5.204007
Log(Country Mass)	6.404508	2.364738	0	10.79853
Cosine Similarity	.099099	.1573169	.001	1
Distance d_1	4.509807	2.544627	0	6.907755
Distance d_2	6.782846	.2363022	3.921989	6.906755
Distance d_3	6.603674	.3943458	3.241714	6.875622
Log(Fidelity Similarity)	-4.392618	2.606018	-6.907755	0

The results for the base models are reported in two separate sections, corresponding to the relationship between the headquarters and the subsidiary (section 4.2) and the relationship between the host country and the subsidiary (section 4.3). All results include firm and year fixed effects and robust clustered standard errors. Figures are rounded to three digits. All of the results tables include a short description of the models, significance levels, and variables. The base model results are reported in Table 3 (headquarters-subsidiary) and Table 4 (host country-subsidiary), which report clusters based on country. Table 5 includes robustness checks for the base gravity model for headquarters-subsidiary, while Table 6 includes robustness checks for the base gravity model for host country-subsidiary. The robustness checks are discussed in section 4.4. Section 4.5 describes the marginal effects and implications for firms as subsidiary sizes change. Section 4.6 presents the results of a mediation mechanism based on immigration from treaty-favored countries. Sections 4.5 and 4.6 follow the theory of section 2.7.

Knowledge Flows between Subsidiary and Headquarters

We find that baseline hypothesis *0a*, which corresponds to a gravitational model for the relationship between an MNC's headquarters and its subsidiary, is validated. All coefficients corresponding to the independent variables are significant and of the expected sign (positive and significant coefficients for masses and negative and significant coefficients for the three distance models, labeled **models 2, 3, and 4**). Additionally, we introduce a gravitational-like model with two similarity measures, cosine similarity (**model 1**) and fidelity similarity (**model 5**), which measure affinity in interests between pairs of (headquarters, subsidiary). These similarity measures are also highly significant, and, as expected, have positive coefficients. The results from models 1-5, corresponding to the subsidiary to headquarters relationship, are shown in Table 3. We observed a stronger effect of the mass of the subsidiary (three to five times greater, depending on

the model) as compared to the effect of the mass of the headquarters on the measure of knowledge flow from subsidiary to headquarters (the coefficient for the logged headquarters mass ranges from 0.29 to 0.35, whereas the coefficient for the subsidiary mass ranges from 0.67 to 0.81). The closest result to a true gravity relationship is that of the fidelity similarity (column 5), where the coefficient approaches 2. Recall that in classical physics, gravity is modeled as inversely proportional to the square of the distance, a relationship most closely approximated by the fidelity similarity. The cosine similarity is also a good candidate, as the coefficient is between 1 and 2. The similarity measures are highly significant. While cosine similarity is known to the management literature (Younge & Kuhn, 2015), these results show that the fidelity similarity may also be a good candidate for future research on knowledge flows and innovation.

Our panel dataset consists mostly of countries that are small in terms of patenting output. Consequently, the panel dataset contains a large number of zeroes for the mass of the subsidiary, which yields zeroes in the citations-dependent variable for the same values of *i, j,* and *t* (as a subsidiary with zero patenting activity in a given year does not produce any citations of the headquarters). Considering the literature on specifications for gravity models (Silva & Tenreiro, 2006) and our data, we found a zero-inflated regression model to be most appropriate for this problem. Our dependent variable is a count variable; the most appropriate model was a zero-inflated negative binomial model. We tested binomial versus zero-inflated negative binomial models and found the latter more appropriate.

Table 3: Subsidiary to Headquarters Regressions - Base Model and Additional Specifications Based on Knowledge Distance

	(1)	(2)	(3)	(4)	(5)
Log(Mass HQ)	0.294^{+}	0.317 $^{+}$	0.294^{+}	0.317^{+}	0.325^{+}
	(0.178)	(0.183)	(0.178)	(0.183)	(.178)
Log(Mass Subsidiary)	0.801***	0.811 ***	0.801***	0.811***	0.679***
	(0.064)	(0.131)	(0.064)	(0.131)	(0.068)
Cosine Similarity	1.301***				
	(0.247)				
Distance d_1		-0.245			
		(0.184)			
Distance d_2			-0.0013***		
			(0.0002)		
Distance d_3				-0.245	
				(0.184)	
Fidelity Similarity					2.079**
					(.392)
_cons	-7.450***	-6.601***	-6.149***	-6.601***	-7.578***
	(1.188)	(1.788)	(1.245)	(1.788)	(1.154)
Year FE	yes	yes	yes	yes	yes
Firm FE	yes	yes	yes	yes	yes
inflate					
log_mass_subsidiary	-4.316**	-2.373	-4.316**	-2.374	-4.328**
	(1.643)	(3.000)	(1.643)	(3.000)	(-3.01)
_cons	3.655***	2.377	3.656***	2.377	3.660***
	(0.681)	(1.865)	(0.681)	(1.864)	(0.622)
lnalpha	-0.354	-0.459	-0.355	-0.321	-0.321
N	3157	3157	3157	3157	3157
Clusters	69	69	69	69	69

Standard Errors in parentheses (country level clustering)
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The first model (1) represents the base model for the headquarters to subsidiary relationship. The dependent variable is the count of citations from the subsidiary patents to the headquarter patents. Main model variables represent stock of knowledge at headquarters as represented by patenting output (Mass HQ), stock of knowledge at subsidiary (Mass Subsidiary), and the cosine similarity between knowledge production at headquarters and subsidiary. The following three models, (2) through (4) use the same base specification, except the cosine similarity is replaced by different measures of knowledge distance between subsidiary and headquarters (as described in the paper). The fifth model (column (5)) utilizes the same base specification, except that the cosine similarity is replaced by the fidelity similarity. All five models include firm and year fixed fixed effects as well as robust clustered standard errors (clustered based on country, 69 clusters). Specifications including controls for the Headquarters to Subsidiary base model are in Table 5.

We excluded three companies with minimal or no international patenting activity from our sample. Specifically, Amazon.com and VERIZON exhibited virtually no foreign subsidiary patents filed with the USPTO during our sample period (2005–2011). The third company removed was AT&T, which has a very complex set of LLCs set up to hold and obfuscate its IP ownership. Because of the low visibility of its IP activities, we were unable to obtain a complete dataset pertaining to AT&T and had to remove it from our regressions. These considerations were also applied for the host country to subsidiary regressions. We ran over 25 additional regressions in robustness checks to validate our findings; we describe these efforts in section 4.4. The results of the regression for subsidiary to host country are shown in Table 4 and detailed in the next section.

Knowledge Flows between Subsidiary and Host Country

We find that baseline hypothesis *0b*, corresponding to a gravitational model for the relationship between an MNC subsidiary and its host country, is validated. The results for this section are reported in Table 4. All coefficients corresponding to the independent variables are significant and of the expected sign (positive and significant coefficients for masses and negative and significant coefficients for the three distance models, labeled **models 2, 3, and 4**). Unlike the case of the subsidiary to headquarters relationship, in the subsidiary-host country relationship, we observe a much stronger effect of the country mass on the knowledge flow as compared to the effect of the subsidiary mass. In the host country-subsidiary relationship, we find the impact of the mass of the host country to be about the same as that of the mass of the subsidiary (the coefficients for both are in the 0.7–0.8 range). This suggests that the relative impact of the host country on the patenting of an R&D subsidiary is stronger than that of the subsidiary's headquarters. We explore this finding further in the marginal effects discussion of section 4.5.

Table 4: Subsidiary to Host Country Regressions Base Model and Additional Specifications Based on Knowledge Distance

	(1)	(2)	(3)	(4)	(5)
Log(Country Mass)	0.757***	0.736***	0.762***	0.759***	0.767***
	(0.450)	(0.039)	(0.047)	(0.046)	(0.048)
Log(Subsidiary Mass)	0.782***	0.816***	0.805***	0.796***	0.675***
	(0.091)	(0.096)	(0.089)	(0.090)	(0.118)
Cosine Similarity	1.306**				
	(0.548)				
Distance d_1		-0.125*			
		(0.109)			
Distance d_2			-0.555$^+$		
			(0.030)		
Distance d_3				-0.441**	
				(0.226)	
Fidelity Similarity					2.357*
					(1.04)
_cons	-13.043***	-12.403***	-9.188***	-10.042***	-13.130***
	(0.681)	(0.685)	(2.116)	(1.610)	(0.733)
Year FE	yes	yes	yes	yes	yes
Firm FE	yes	yes	yes	yes	yes
inflate					
Log(Mass Subsidiary)	-2.080***	-1.907***	-2.120***	–2.081***	-2.100***
	(0.223)	(0.283)	(0.209)	(0.215)	(0.231)
_cons	2.479***	2.301***	2.549***	2.492***	2.500***
	(0.345)	(0.443)	(0.332)	(0.348)	(0.347)
lnalpha	-0.294	-0.288	-0.283	-0.285	-0.284
N	3157	3157	3153	3153	3157
Clusters	69	69	69	69	69

Standard Errors in parentheses (country clusters)
$^+$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The first model (1) represents the base model for the host country to subsidiary relationship. The dependent variable is the count of citations from the subsidiary patents to the host country patents. Main model variables represent stock of knowledge of the host country as represented by patenting output (Country Mass), stock of knowledge at subsidiary (Mass Subsidiary), and the cosine similarity between knowledge production in the host country and that of the subsidiary. Models (2) through (4) use the same base specification, except the cosine similarity is replaced by different measures of knowledge distance between subsidiary and country (as described in the paper). The fifth model (column (5)) utilizes the same base specification, except that the cosine similarity is replaced by the fidelity similarity. All five models include firm and year fixed fixed effects as well as robust clustered standard errors (clustered based on country, 69 clusters). Specifications including controls for the Country to Subsidiary base model are in Table 6.

We also notice that some of the distance/similarity measures perform slightly worse in terms of significance than the same measures in the headquarters-subsidiary case. This is intuitive – the innovation output on average for the countries is much larger than that of the company headquarters, and our knowledge similarity and distance measures are all based on comparing the spectrum of innovation of the two parties (headquarters-subsidiary or subsidiary-host country). If the host country is far larger than the headquarters, it produces diverse innovation and an innovation spectrum that is far noisier than a company subsidiary spectrum; the similarity in such a case is fairly poor. All similarity/distance measures remain significant and of the expected sign. The coefficient on fidelity similarity is closer to 2 than that of the cosine similarity, again suggesting that it is a better fit for a true gravity relationship.

Robustness Checks

In addition to the ten models in Tables 2 and 3 corresponding to the five measures of knowledge distance applied to each setting (headquarters-subsidiary and host country-subsidiary), an additional 27 regressions were run to test the robustness of the knowledge flow gravity model. Control variables were obtained from a diverse array of sources, ranging from US State Department visa data to UN immigration data, SCOPUS-based measurements of scientific output, Organisation de Coopération et Devéloppement Economiques (OECD) country-level controls for employment and educational achievement, and the World Bank. The direction and overall magnitude of the main model variables (mass of headquarters, mass of subsidiary, mass of host country, and cosine similarity) were maintained throughout these checks.

The first set of checks, in Table 5, corresponds to four models obtained by adding controls to the base model for the headquarters-subsidiary relationship. Because of the large number of countries (66) in our dataset and the different coverage of those countries in our various data sources, running

all of the control variables in one model would result in a vanishingly small subset of our data. To mitigate this problem, we chose to run groups of control variables from the same source within the same model and to split the control variables among four different models. The first column of Table 5 represents the base model for the headquarters-subsidiary relationship (main model independent variables: mass headquarters, mass subsidiary, cosine similarity, with firm and year fixed effects) and is used as a comparison for the next four columns. Model (2) in Table 5 corresponds to the base headquarters-subsidiary model and includes the controls sourced from the OECD.[8] We used two measures of the potential for R&D among a country's population: percent of population attaining tertiary education levels and percent of population attaining a PhD. We chose these measures because highly skilled labor may affect an MNC's decision to locate an R&D subsidiary in a given country and its decision to hire locally, which may in turn affect the knowledge flows to the subsidiary from the headquarters or the host. Yearly data were available for tertiary education; data for PhD graduates, however, were sparser, so we estimated the yearly values using CAGR. These two variables were run as part of our "OECD Controls" category and did not affect the results for our main variables; furthermore, PhDs as a percentage of the population were not significant.

SCOPUS is owned by Elsevier and marketed as "the largest database of peer-reviewed literature."[9] It can be used to derive measures of scientific output and quality at the country level. SCImago Journal & Country Rank is a database of country-level measures derived from SCOPUS and made available by SCImago Lab (in partnership with Elsevier).[10] Scientific output (number of articles) and quality (H-index) are measured at the country-year level and are made available by SCImago.

[8] OECD Research and Development Indicators; see for example https://data.oecd.org/rd/researchers.htm#indicator-chart, Accessed 1/1/2017.
[9] SCOPUS, https://www.elsevier.com/solutions/scopus. Accessed 1/1/2017.
[10] SCImago Journal & Country Rank, http://www.scimagojr.com/countryrank.php. Accessed 1/1/2017.

We used the country-level variables as controls in Model (3) of Table 5. The overall magnitude of the coefficients of the main variables and sign did not change. Neither control variable was significant.

The US State Department, as a taxpayer-funded agency, makes data of public interest freely available. The Non-Immigrant Visa count per country-year is a relevant source for this study, as certain visa categories are tied to specific types of economic activity. We looked at the following visa categories: B1, one of the most common visa types, which allows for non-immigrant business travel (short stays, such as brief collaborations or conferences); H1B, the most-used skilled worker immigration visa; L1, specifically for intra-company transferees from outside the US to the US and tailored to employees transferring within MNCs; J1, tailored to academic exchanges and used by teachers, scholars, students, and specialists; and O1, extraordinary ability visas reserved for the most desirable specialists and researchers. These types of visas all favor economic exchange between the US and another country and so may be relevant to the headquarters-subsidiary relationship. None of the visa-based variables were significant, and the overall magnitudes and signs of our main coefficients did not change (Model 4, Table 5). Similarly, we used overall immigration counts to the US from other countries as found in UN data[11] (log of number of immigrants per country-year) and found that this variable was not significant. Moreover, the relative magnitudes and directions of the coefficients for the main variables did not change.

We also tested our results using a control for the number of researchers per million inhabitants (sourced from the World Bank) as another measure of a country's potential for R&D. This control variable was not significant and is not reported in Table 5 due to space considerations. Additional

[11] UN Population Division compiles migration flows data at the country-year level, http://www.un.org/en/development/desa/population/migration/data/empirical2/migrationflows.shtml. Accessed 1/1/2017.

robustness checks for the headquarters-subsidiary side include clustering standard errors based on firms (no change, all five models) and, to verify that our effects are not driven by a few countries that are very prolific in terms of patenting (such as the UK, India, China, and France), we ran the five models using small patenting countries only. Again, we observed no change in the results. The latter are not reported here because of space limitations.

The checks for the host country-subsidiary relationships are presented in Table 6. As in Table 5, the base model is the first column in Table 6, serving as a reference for the other models. The first set of controls, in column (2) of Table 6, represents the base gravity model of the host country-subsidiary but includes the OECD controls described previously. The third column (Model (3)) of Table 6 includes the SCOPUS-based controls; the fourth column represents results that include the World Bank-sourced number of researchers per million; the fifth column shows results that include the Immigration to US from host country variable (sourced from the UN); the sixth column of Table 6 includes the State Department Visa issuance counts for categories related to business/worker exchange/skilled labor (B1, J1, L1, H1B, O1).

As in Table 5, all results in Table 6 are based on models that include firm and year fixed effects and clustered robust standard errors (country-level clusters). The magnitudes of the main coefficients and the directions of the effects did not change. In addition to the results reported in Table 6, we ran all five models from Table 4 with firm-level clusters. The results (not reported here due to space limitations) did not change.

Table 6: Robustness Checks - Subsidiary to Host Country Regressions Subsidiary to Headquarters Regressions - Base Model in Comparison with Models with Various Controls

	(1)	(2)	(3)	(4)	(5)	(6)
Log(Cntry Mass)	0.757***	0.772***	0.844***	0.757***	0.710***	0.831***
	(0.450)	(0.045)	(0.053)	(0.043)	(0.041)	(0.214)
Log(Sub. Mass)	0.782***	0.905***	0.963***	0.977***	0.901***	0.816***
	(0.091)	(0.058)	(0.064)	(0.087)	(0.058)	(0.816)
Cosine Sim.	1.306**	1.076+	0.499	1.519*	0.798	1.332*
	(0.548)	(0.643)	(0.572)	(0.446)	(0.588)	(0.582)
_cons	-13.043***	-17.668***	-11.042***	-14.601***	-13.108***	-13.435***
	(0.681)	(1.076)	(1.507)	(1.114)	(0.599)	(0.903)
Year FE	yes	yes	yes	yes	yes	yes
Firm FE	yes	yes	yes	yes	yes	yes
OECD Ctrls		yes				
SCOPUS Ctrls			yes			
Res per Mln				yes		
Imm to US (UN)					yes	
Visas (State)						yes
inflate						
Log(Subs Mass)	-2.080***	-2.106***	-2.096***	-1.979***	-2.149***	-2.096***
	(0.223)	(0.302)	(0.270)	(0.251)	(0.269)	(0.265)
_cons	2.479***	2.142***	2.367***	2.247***	2.419***	2.410***
	(0.345)	(0.357)	(0.401)	(0.314)	(0.417)	(0.341)
lnalpha	-0.294	-0.571	-0.482	-1.116	-0.409	-0.337
N	3157	1998	3074	2397	3080	2750
Clusters	69	26	68	55	68	56

Standard Errors in parentheses (country clusters)
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The first model (1) represents the base model for the subsidiary to host country relationship. The following four models include controls added to the base specification model. The control variables were obtained from different sources with varying coverage on countries, thus groups of control variables from the same source are displayed in different columns as part of separate regressions. Model (2) represents the base specification plus OECD controls measuring research characteristics at the country level. Model (3) represents the base specification plus controls based on SCOPUS data (source: SCImago Country level data) which measure country level scientific throughput. Model (4) is the base specification plus a control based on World Bank country level data on researchers per million. Model (5) is the base specification plus a control variable for overall immigration to the US based on country of origin (data sourced from the UN). Model (6) is the base specification including US State Dept Visas per country for work/exchange categories. Additional robustness checks were run and are discussed in the paper. All results in this table include firm and year fixed effects as well as robust clustered standard errors (clustered based on country).

Comparison of Marginal Effects

The absorptive capacity of the MNC subsidiary for locally sourced knowledge grows with its existing knowledge stock. However, HQ-sourced knowledge may no longer be as relevant to a growing subsidiary potentially seeking a degree of autonomy from the MNC. To test our hypothesis regarding the effects of a growing knowledge stock at the subsidiary on its sources of knowledge flows, we ran marginal effects in the cosine similarity model for both the subsidiary-headquarters and subsidiary-host country relationships, based on variations in the size of the subsidiary (logged mass of subsidiary).

The results, reported in Figure 2, show that as the size of the subsidiary increases, the influence of the host country on the innovation of the subsidiary grows at a faster rate than the influence of the headquarters. This effect is especially visible for the largest subsidiaries (mean + 2 SD) in our sample. For example, for a firm close to the mean HQ size, a one standard deviation increase in the knowledge stock ("mass") of the subsidiary would yield about double the growth of patent citations to the subsidiary's host country as compared to its headquarters. We find possible heterogeneity across MNC subsidiaries in the relative importance of the host country context and the headquarters. Our results indicate that the host country context is more important for subsidiaries with a growing stock of patented innovations.

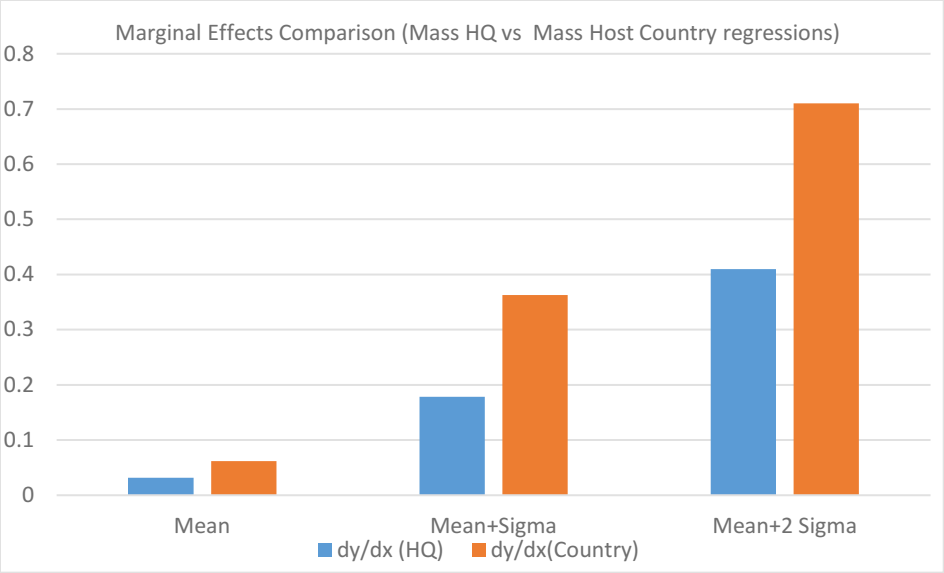

Figure 2. Marginal effects of mass headquarters and mass host country at various subsidiary sizes.

Cisco Case Study. In 1995, CISCO extended its operations in India. As a source of highly-skilled engineers and a country that is welcoming to foreign MNCs, India has been a host of major R&D subsidiaries for many of the top-patenting MNCs headquartered in the US, our target sample. While many technology firms did create subsidiaries in India, CISCO went a step further, creating a second R&D HQ in India in 2006 (Kapuri, 2006) – its "largest global development center outside the US" that "develops disruptive business models for Cisco to create new go-to-market channels, markets, processes and technologies for emerging markets."[12] This prestigious designation yielded a significant investment in the development of the center and a degree of independence as one of the top US HQ-based CISCO executives moved to head the new Global Development Center in Bangalore (Kapuri, 2006). Based on our approach, we expected this event to yield a stronger tie to local context innovation and a continued interest in commercializing the technologies developed

[12] CISCO Company Overview, https://www.cisco.com/c/en_in/about/company-overview.html. Accessed 3/26/2018.

at the India Global Development Center in the US markets. Our gravity model predicts that the distance between CISCO India post-2006 and its host country patenting should shrink. Figure 3 shows the cosine similarity between the patenting in CISCO India and the host country patenting and compares it to the average cosine similarity of all firms in our sample to patenting in India: there is a jump in similarity between CISCO India and its host country post-2006 and a steeper slope in the increase of the similarity as compared to the average Indian subsidiary in our sample.

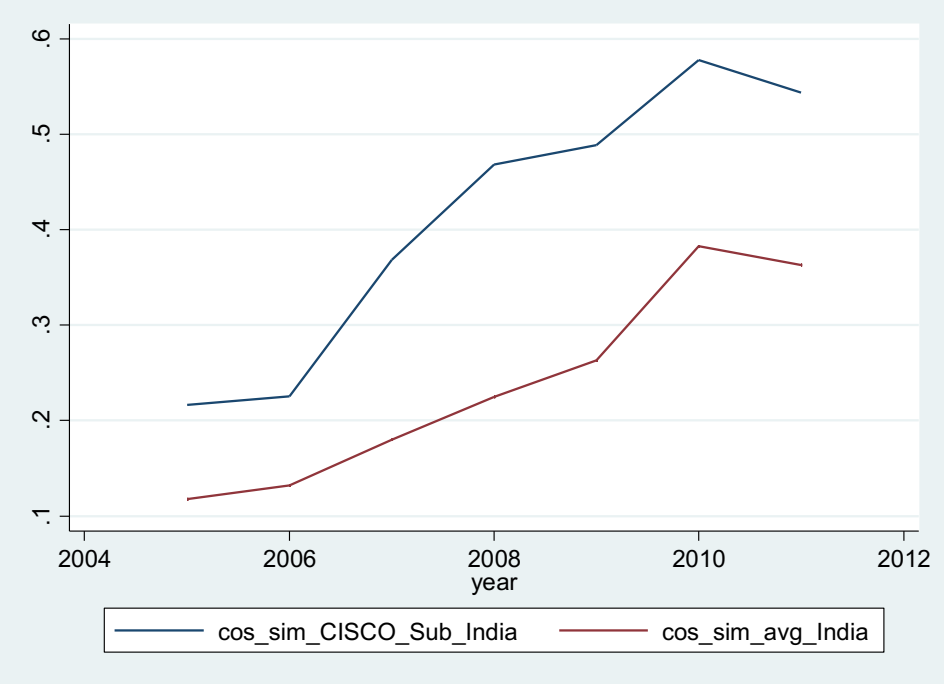

Figure 3. Comparison of cosine similarity between CISCO's India subsidiary patents and host country patents to the average cosine similarity between any Indian subsidiary filed patents and host patents.

This case study, while limited in scope to the one firm in our sample that had a second HQ designation, lends additional support to our findings and serves as an example of where a gravity model of knowledge flows could be used to predict future strategic investments abroad by MNCs.

Mediation of Headquarters – Subsidiary Relationship through Immigration

As Björkman *et al*. (2004) and O'Donnell (2000) found, MNCs can use managerial socialization as a mechanism to bring subsidiaries closer to the HQ, including through travel and training programs (Björkman *et al*., 2004). As all firms in our dataset are US-headquartered, we reviewed the various US visa categories to determine which types of visa enable such interactions. We particularly looked at high-skill short-term employment visas that enable short-term exchanges between company units. We expected this short analysis to shed light on use of the E-1 visa.

The US has established a special category of visa for countries it deems key trading partners called the E-1 Treaty Trader Visa, for "essential employees, employed in a supervisory or executive capacity, or possess[ing] highly specialized skills essential to the efficient operation of the firm", per the State Department E-1 Visa applicant guidelines.[13] Only certain countries are designated as essential trading partners. Some of the trade treaties have been in place for over a century[14] and signify substantial economic trade and collaboration between the US and that country, with the E-1 visa constituting one component of that bilateral trade. This type of visa is restricted to the most highly-skilled workers; ordinary skilled or unskilled workers do not qualify. This particular type of visa is of interest to our study because it measures only the flow of key employees, either

[13] US State Department E-1 Application, https://travel.state.gov/content/visas/en/employment/treaty.html. Accessed 1/2/2017.
[14] US State Department, List of Trade Treaty Countries, https://travel.state.gov/content/visas/en/fees/treaty.html. Accessed 1/2/2017.

managers or specialists, from the treaty country to the US. The E-1 is a nonimmigrant visa, which restricts the amount of time these employees can spend in the US while remaining employed at the same firm. These highly influential employees return to their home countries with the knowledge and collaborations they have forged in the US. We expect this flow to influence interest in innovating for the US market and consequently to result in an increased number of US patents filed by citizens of the treaty countries. The relationship between the mass of the country and the subsidiary's citations of the host country's patents may be mediated by the number of essential employees granted E1 visas. To test this expectation, we ran a structural mediation model and found that E1 visa counts account for approximately 16% of the total effect of country mass (Table 7). This result may be relevant to policymakers, as encouraging an increase in E1 visas may increase the number of patents filed with the USPTO and so benefit the US economy. It may also be of interest to managers of MNCs as an additional mechanism to strengthen relationships with subsidiary management.

Table 7: Mediator Variable Treaty Trader Visas (E1 category) - Host Country to Subsidiary Relationship

	(E1)	Robust Sobel	Robust Goodman
Indirect Effects			
Log(Country Mass)	0.049*	yes	yes
	(0.025)		
Direct Effects			
Log(Country Mass)	0.252***		
	(0.037)		
Total Effects			
Log(Country Mass)	0.301***		
	(0.028)		

Standard Errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Trade Treaty Visas (E1 category) act as a mediator for the effect of the country mass on the count of citations from the subsidiary to its host country. Treaty Trader Visa numbers are from the U.S. State Department. Sobel and Goodman tests were passed at a significance level of 0.05.

The proportion of the total effect of country mass mediated by log(E1) is: 16.228%

CONCLUSIONS AND DISCUSSION

Summary of Results

In this paper, we use the gravity equation in economics to estimate two comparable specifications – one measuring knowledge flows from the headquarters to a multinational subsidiary and the second measuring knowledge flows from the host country context to the subsidiary. Our empirical apparatus allows us to consider a relatively robust "apples to apples" comparison of these two knowledge flows. Using unique data on patent citations and the "masses" of patents filed with the USPTO by the headquarters, subsidiaries and host countries of the top 25 US-headquartered MNCs from 2005 to 2011, we validate the gravity specification for both headquarters to subsidiary knowledge flows and host country to subsidiary knowledge flows. Our results indicate that the role of the subsidiary mass is several times more important than the role of the headquarters mass.

Moreover, the influence of the subsidiary mass on the knowledge flow is increasing faster than proportionally. This is a departure from the standard gravitational model.

Importance of the Host Country Context for MNC Knowledge Flows

We can also compare the *relative influence* of the headquarters and the host country context on knowledge flows into the MNC subsidiary. Our results indicate that as the size of the subsidiary increases, the host country's influence on knowledge flows into the subsidiary grows faster than the influence of the headquarters. In other words, MNC subsidiaries may differ in the extent to which they are influenced by the host country as compared to the headquarters. Specifically, subsidiaries that have a greater stock of patented innovations may be more susceptible to the host country's influence.

Our results respond to the recent call in the strategy and international business literature for firms to develop contextual intelligence (Dhanaraj & Khanna, 2011; Khanna, 2015). Meyer, Mudambi, and Narula (2011) predict that host countries will play an increasingly important role in shaping MNC subsidiaries and thus MNCs overall. Santos and Williamson (2015) advise MNCs to cultivate a local presence that is not merely "adaptive" but fully intertwined with or even "made" in the local context. One way to establish a local presence is by learning from the host country context. In summary, our results represent a step forward in empirically measuring reverse innovation (Govindrajan & Ramamurti, 2011) and comparing knowledge flows from the host country to the subsidiary.

The results are robust to the removal of the large patenting countries, again suggesting that the results are driven by the smaller patenting countries. Given our observations from the marginal effects analysis, firms should acknowledge that locating a subsidiary in a country with low

patenting activity may have a different impact on that subsidiary's R&D trajectory than that produced by locating it in a country with high patenting activity, and that the subsidiary's direction may become more independent of the headquarters as it grows.

Unique Measures of Knowledge Distance

We contribute to the knowledge flow and gravity literatures by introducing unique measures of knowledge distance. We depart from the traditional gravity model that uses a physical distance measure and instead propose a measurement of knowledge distance based on cosine similarity and several transformations of the cosine similarity, and we introduce fidelity similarity (the Bhattacharya coefficient) to the management literature. We find the fidelity similarity to be the truest to the classical physics gravity model and propose it for future use in our literature. The similarity approach presented here is useful for comparing the patenting outputs of entities in general and is not limited to MNCs; in other words, we expect this part of our study to be generalizable.

Limitations and Future Research Directions

We are aware that our study is limited by our use of only US-headquartered MNCs and USPTO data. Even with this limitation, obtaining data of high enough quality to use was a herculean task, given the number of errors in raw patent data that we had to manually and programmatically correct.[15] Many errors in firm names had to be manually corrected. Future studies should examine the influences of the communication and control forms practiced by MNC headquarters, as identified by Nobel and Birkinshaw (1998), on the knowledge flow among subsidiaries, the headquarters, and the host countries. Finally, a detailed mathematical analysis of the distances used

[15] We are aware that there are other excellent databases, but we needed extra information for our study.

in this study might explain why, while some of them are isomorphic (d_1, d_2, d_3), they are considerably different in the extent to which they reveal differences in innovative processes between entities. We suggest that the methodology of evaluating distances by the spectrum of the interest in the innovative domains, as proposed in this study and found outstandingly effective, could be applicable to a broader class of economic statistical models. In future work, this methodology deserves to be tested in various other problems of knowledge transfer.

REFERENCES

Agrawal, A., Cockburn, I., & McHale, J., 2006. Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. ***Journal of Economic Geography***, 6(5): 571-591.

Agrawal, A., Kapur, D., & McHale, J., 2008. How do spatial and social proximity influence knowledge flows? Evidence from patent data. ***Journal of Urban Economics***, 64(2): 258-269.

Alcacer, J., & Zhao, M. 2012. Local R&D strategies and multilocation firms: The role of internal linkages. ***Management Science,*** 58(4): 734–753.

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. 2004. Subsidiaries and knowledge creation: The influence of the MNC and host country on innovation. ***Strategic Management Journal,*** 25(8-9): 847 –864.

Anderson, J.E., 1979. A theoretical foundation for the gravity equation. ***The American Economic Review***, 69(1): 106-116.

Andersson, U., Forsgren, M., & Holm, U. 2002. The strategic impact of external networks: Subsidiary performance and competence development in the multinational corporation. ***Strategic Management Journal,*** 23(11): 979–996.

Anderson, J.E., 1979. A theoretical foundation for the gravity equation. ***The American Economic Review***, 69(1): 106-116.

Anderson, J. E., & van Wincoop, E. 2003. Gravity with gravitas: A solution to the border puzzle. ***American Economic Review,*** 93(1): 170–192.

Baier, S.L. and Bergstrand, J.H., 2009. Estimating the effects of free trade agreements on international trade flows using matching econometrics. ***Journal of International Economics***, 77(1): 63-76.

Bartlett, C. A. 1986. Building and managing the transnational: The new organizational challenge**.** In M. Porter (Ed.), ***Competition in global industries***, pp. 367-401. Boston, MA: Harvard Business School Press.

Bartlett, C. A., Doz, Y., & Hedlund, G. 2012. ***Managing the global firm*** (RLE International Business). Vol. 3. London: Routledge.

Bergstrand, J.H., 1985. The gravity equation in international trade: Some microeconomic foundations and empirical evidence. ***The Review of Economics and Statistics***, 67(3): 474-481.

Bergstrand, J. H. 1989. The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade. ***The Review of Economics and Statistics,*** 71(1): 143–153.

Birkinshaw, J., & Hood, N. 1998. Multinational subsidiary evolution: Capability and charter change in foreign-owned subsidiary companies. ***Academy of Management Review***, 23(4): 773-795.

Birkinshaw, J. M., & Morrison, A. J. 1995. Configurations of strategy and structure in subsidiaries of multinational corporations. ***Journal of International Business Studies,*** 26(4): 729–753.

Björkman, I., Barner-Rasmussen, W., & Li, L. 2004. Managing knowledge transfer in MNCs: The impact of headquarters control mechanisms. ***Journal of International Business Studies,*** 35(5): 443–455.

Buckley, P. J., & Casson, M. 1976. ***The future of the multinational enterprise***. Vol. 1. London: MacMillan.

Caves, R. E. 1971. International corporations: The industrial economics of foreign investment. ***Economica,*** 38(149): 1–27.

Caves, R. E. 1974. Multinational firms, competition, and productivity in host-country markets. ***Economica,*** 41(162): 176–193.

Caves, D. W., Christensen, L.R., & Diewert, W. E. 1982. The economic theory of index numbers and the measurement of input, output, and productivity. ***Econometrica: Journal of the Econometric Society,*** 50(6): 1393–1414.

Choudhury, P. 2016. 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.

Cohen, W. M., & Levinthal, D. A. 1990. Absorptive capacity: A new perspective on learning and innovation. ***Administrative Science Quarterly,*** 35(1): 128-152.

Dacin, M. T., Beal, B. D., & Ventresca, M. J. 1999. The embeddedness of organizations: Dialogue & directions. **Journal of Management**, 25(3): 317–356.

Deza, M. M., & Deza, E. 2015. ***Encyclopedia of distances***. Berlin: Springer.

Doz, Y. L., Bartlett, C. A., & Prahalad, C. K. 1981. Global competitive pressures and host country demands managing tensions in MNCs. ***California Management Review***, 23(3): 63–74.

Dhanaraj, C., and Khanna, T. 2011. Transforming mental models on emerging markets. ***Academy***

of Management Learning & Education, 10(4): 684–701.

Feinberg, S. E., & Gupta, A. K. 2004. Knowledge spillovers and the assignment of R&D responsibilities to foreign subsidiaries. ***Strategic Management Journal***, 25: 823–845.

Ghoshal, S. 1986. ***The innovative multinational: A differentiated network of organizational roles and management processes***. Unpublished thesis, Harvard University, Graduate School of Business Administration.

Ghoshal, S. 1987. Global strategy: An organizing framework. ***Strategic Management Journal***, 8(5): 425–440.

Ghoshal, S., & Bartlett, C. A. 1990. The multinational corporation as an interorganizational network. ***Academy of Management Review***, 15(4): 603–626.

Ghoshal, S., & Nohria, N. 1989. Internal differentiation within multinational corporations. ***Strategic Management Journal,*** 10(4): 323–337.

Govindarajan, V. & Ramamurti, R., 2011. Reverse innovation, emerging markets, and global strategy. ***Global Strategy Journal***, 1(3-4) : 191-205.

Gupta, A. K., & Govindarajan, V. 1991. Knowledge flows and the structure of control within multinational corporations. ***Academy of Management Review***, 16(4): 768–792.

Gupta, A. K., & Govindarajan, V. 2000. Knowledge flows within multinational corporations. ***Strategic Management Journal,*** 21(4): 473–496.

Hansen, M. T. 2002. Knowledge networks: Explaining effective knowledge sharing in multiunit companies. ***Organization Science***, 13(3): 232–248.

Hymer, S.H., 1960. ***The International Operations of National Firms: A Study of Direct Investment.*** Unpublished Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge.

Hymer, S. H. 1976. ***The international operations of national firms: A study of direct foreign investment.*** Vol. 14. Cambridge, MA: MIT Press.

Jaffe, A.B., Trajtenberg, M. and Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. ***Quarterly Journal of Economics***, 108(3): 577-598.

Kapur, M. 2006. Cisco to set up second head office in Bangalore. *The Times of India* (Nov 2), https://timesofindia.indiatimes.com/business/india-business/Cisco-to-set-up-second-head-office-in-Blore/articleshow/287171.cms. Accessed April 16th, 2018.

Kay, L., Newman, N., Youtie, J., Porter, A.L., & Rafols, I. 2014. Patent overlay mapping: Visualizing technological distance. ***Journal of the Association for Information Science and Technology,*** 65(12): 2432–2443.

Khanna, T. A case for contextual intelligence. 2015. ***Management International Review,*** 55(2): 181–190.

Kindleberger, C. P. 1969. American business abroad. ***The International Executive,*** 11(2): 11–12.

Kostova, T., & Roth, K. 2002. Adoption of an organizational practice by subsidiaries of multinational corporations: Institutional and relational effects. ***Academy of Management Journal,*** 45(1): 215–233.

Lewer, J. J., & Van den Berg, H. 2008. A gravity model of immigration. ***Economics Letters,*** 99(1): 164–167.

Mátyás, L. 1997. Proper econometric specification of the gravity model. ***The World Economy***, 20(3): 363–368.

Meyer, K. E., Mudambi, R., & Narula, R. 2011. Multinational enterprises and local contexts: The opportunities and challenges of multiple embeddedness. ***Journal of Management Studies,*** 48(2): 235–252.

Monteiro, L. F., Arvidsson, N., & Birkinshaw, J. 2008. Knowledge flows within multinational corporations: Explaining subsidiary isolation and its performance implications. ***Organization Science***, 19(1): 90–107.

Mudambi, R., Pedersen, T., & Andersson, U. 2014. How subsidiaries gain power in multinational corporations. ***Journal of World Business***, 49(1): 101–113.

Narula, R. 2014. Exploring the paradox of competence-creating subsidiaries: Balancing bandwidth and dispersion in MNEs. ***Long Range Planning***, 47(1): 4–15.

Nell, P. C., & Ambos, B. 2013. Parenting advantage in the MNC: An embeddedness perspective on the value added by headquarters. ***Strategic Management Journal***, 34(9): 1086–1103.

Nelson, R., & Winter, S. 1982. ***An evolutionary theory of economic change***. Cambridge, MA: Belknap Press of Harvard University Press.

Nobel, R., & Birkinshaw, J. 1998. Innovation in multinational corporations: Control and communication patterns in international R&D firms. ***Strategic Management Journal***, 19: 479–496.

Nohria, N., & Ghoshal, S. 1994. Differentiated fit and shared values: Alternatives for managing headquarters-subsidiary relations. ***Strategic Management Journal***, 15(6): 491–502.

Oettl, A., & Agrawal, A., 2008. International labor mobility and knowledge flow externalities. ***Journal of International Business Studies***, 39(8): 1242-1260.

O'Donnell, S. W. 2000. Managing foreign subsidiaries: Agents of headquarters, or an interdependent network? ***Strategic Management Journal,*** 21(21): 525–548.

Porter, M. E. 1986. ***Competition in global industries***. Boston, MA: Harvard Business Press.

Rugman, A. M., & Verbeke, A. 1992. A note on the transnational solution and the transaction cost theory of multinational strategic management. ***Journal of International Business Studies,*** 23(4): 761–771.

Rugman, A., & Verbeke, A. 2001. Subsidiary-specific advantages in multinational enterprises. ***Strategic Management Journal***, 22(3): 237–250.

Santos, J.F., & Williamson, P.J., 2015. The new mission for multinationals. ***MIT Sloan Management Review***, 56(4): 45.

Silva, J. M. C. S., & Tenreyro, S. 2006. The log of gravity. ***The Review of Economics and Statistics,*** 88(4): 641–658.

Singh, J. 2007. Asymmetry of knowledge spillovers between MNCs and host country firms. ***Journal of International Business Studies***, 38(5): 764–786.

Singh, J. 2008. Distributed R&D, cross-regional knowledge integration and quality of innovative output. ***Research Policy***, 37(1): 77–96.

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.

Stahl, G. K., Tung, R. L., Kostova, T., & Zellmer-Bruhn, M. 2016. Widening the lens: Rethinking distance, diversity, and foreignness in international business research through positive organizational scholarship. ***Journal of International Business Studies,*** 47(6): 621–630.

Summary, R.M., 1989. A political-economic model of US bilateral trade. ***Review of Economics and Statistics***, 71(1):179-182.

Teece, D. J. 1981. The market for know-how and the efficient international transfer of technology. ***The Annals of the American Academy of Political and Social Science***, 458(1): 81–96.

Vernon, R. 1966. International investment and international trade in the product cycle. ***Quarterly Journal of Economics,*** 80(2): 190–207.

Waugh, M.E., 2010. International trade and income differences. ***American Economic Review***, 100(5): 2093-2124.

Younge, K. A., & Kuhn, J. M. 2015. Patent-to-patent similarity: A vector space model. Available at SSRN.

Zahra, S. A., & George, G. 2002. Absorptive capacity: A review, reconceptualization, and extension. ***Academy of Management Review,*** 27(2): 185–203.

Zhao, M. 2006. Conducting R&D in countries with weak intellectual property rights protection. ***Management Science***, 52(8): 1185–1199.

Chapter 2

Machine Learning Methods for Strategy Research

Mike H.M. Teodorescu

ABSTRACT

Numerous applications of machine learning have gained acceptance in the field of strategy and management research only in the last few years. Established uses span such diverse problems as strategic foreign investments, strategic resource allocation, systemic risk analysis, and customer relationship management. This survey chapter covers natural language processing methods focused on text analytics and machine learning methods with their applications to management research and strategic practice. The methods are presented accessibly, with directly applicable examples from multiple subfields of management science. Additionally, this chapter presents some applications of machine learning to innovation research, specifically topic modeling and corpus analyses of US patents.

INTRODUCTION

During the last few decades, various management disciplines became heavily dependent on machine learning methods and tools. Domains such as marketing (Gans *et al.*, 2017; Struhl, 2015), financial markets (Tetlock, 2007; Tan *et al*., 2007; Bollen *et al.*, 2011), risk management (Hu *et al.*, 2012, Chen *et al.*, 2012), knowledge management (Williams & Lee, 2009; Li *et al.*, 2014; Balsmeier *et al.*, 2016), and logistics (Jordan & Mitchell, 2015), among others, are inconceivable today without the use of vast quantities of data and machine learning tools. Machine learning is the study of methods that make it possible to find patterns in data and the subsequent use of these

patterns to construct predictions and inferences and to make decisions. The purpose of this chapter is to give a survey of machine learning methods and their applications to management, providing the reader with fundamental methodological tools via steps and examples that are accessible and easily reusable. The interested reader will also find targeted references to in-depth methodological content expanding the methods surveyed here and to a set of relevant articles in our management literature that showcase some of these methods. The examples are presented so as to be usable by a broad audience. Given text-based methods' growing use in our field and their partial independence of the other machine learning methods, the first half of the chapter presents and exemplifies textual analysis methods (part of the field of statistical natural language processing) such as term frequency, textual similarity, corpora considerations, and sentiment analysis (Manning & Schütze, 1999). The second part of the chapter covers general machine learning concepts, such as the concept of classification, the decision boundary, training and testing, cross-validation, and other fundamentals. It also exemplifies typical methods in machine learning that extend beyond text, such as decision trees, random forests, k-Nearest-Neighbors, and Naïve Bayes. Each method is presented together with an implementation in an easy-to-use machine learning toolkit[16] that requires no programming background and with a current management literature example or a potential use in the management literature.

For a quick orientation to the main applications and trends of the methods of machine learning in solving important problems in strategy and management, Table 1 summarizes some of these problems and provides a few relevant references.

[16] The mentions throughout this paper of various toolkits and software packages are not an endorsement of these toolkits and software packages. The opinions expressed are solely of the author, and are based on his experience with these toolkits, languages, and packages.

Table 1. ML in strategy and management research

Domain	Problem treated	Method used
Multinationals	Strategy of foreign investments (Debaere *et al.*, 2010; Roth, 1992). Strategy of resource allocation (Williams & Lee, 2009). Strategy of international marketing, foreign market opportunity assessment (Cavusgil *et al.*, 2004; Hu & Liu, 2004; Singh *et al.*, 1996; Punj & Stewart, 1983; Wedel & Kamakura, 2012). Supply chains (Stock *et al.*, 2000). Analysis of strategic leadership and executive innovation (Elenkov & Wright, 2005).	Cluster analysis; K-means validated with Ward's method (Singh *et al.*, 1996; Williams & Lee, 2009) hierarchical clustering (Williams & Lee, 2009; Stock *et al.*, 2000; Singh *et al.*, 1996). Establishing control groups for firms in DID models, using kNN, for establishing models for foreign investments (Debaere *et al.*, 2010). Web mining, NLP (web intelligence) (Lau *et al.*, 2012). Clustering (Elenkov & Wright, 2005).
Corporate Governance	Assessing CEO personality (Gow *et al.*, 2016). Managerial attention/cognition (Nadkarni & Barr, 2008; Eggers and Kaplan, 2009). CEO strategy and acquisitions (Gamache *et al.*, 2015). Gender effects (Lee, 2007), (Kanze *et al.*, 2017).	PCA (Slack *et al.*, 2010 ; Zhu, 2013; Lange *et al.*, 2014). Text frequency analysis (Kanze *et al.*, 2017 ; Gamache *et al.*, 2015 ; Eggers *et al.*, 2017). Text-based clustering (Gow *et al.*, 2016).
Financial Markets	Stock market prediction (Bollen *et al.*, 2011; Tan *et al.*, 2007; Lugmayr, 2013). Investor sentiment analysis (Tetlock, 2007). Legal issues in finance – liabilities (Loughran & McDonald, 2011).	Classification, prediction, NLP, Web analysis (Bollen *et al.*, 2011; Tan *et al.*, 2007). NLP and Web mining (Lugmayr, 2013; Loughran & McDonald, 2011; Tetlock, 2007).
Banking System	Systemic risk, contagious bank failures, system failure prediction (Hu *et al.* 2012; Chen *et al.*, 2012). Bank failure prediction (Tan *et al.*, 2007).	Classification, prediction, network model: "Network Approach to Risk Management (NARM)," "Rank-In-Network Principle," and "Link-Aware

Domain	Problem treated	Method used
		Systemic Estimation of Risks" (Hu *et al.*, 2012). Network-based modeling (Chen *et al.,* 2012).
Credit Prediction	Individuals and corporate credit scoring, credit worthiness prediction, business failure prediction (Liab & Sun, 2011; Sohn & Kim, 2012; Ju & Sohn, 2014; Nikoloc *et al.,* 2013).	Decision tree, SVM (Sohn and Kim, 2012 ; Cubiles-De-La-Vega *et al.*, 2013). Logistic regression (Nikoloc *et al.*, 2013). PCA-based (Liab & Sun, 2011). Neural network (NN) (Liab & Sun, 2011 : Cubiles-De-La-Vega *et al.*, 2013). Classification trees (Cubiles-De-La-Vega *et al.*, 2013) and decision trees (Sohn & Kim, 2012).
Market Segmentation	Market-level analysis, segmentation (Chiu *et al.*, 2009; Punj & Stewart, 1983; Wang, 2009; Wedel & Kamakura, 2012).	k-means, particle swarm optimization (Chiu *et al.*, 2009). Cluster analysis (Punj & Stewart, 1983), kernel-based clustering (Wang, 2009), various clustering techniques (Wedel & Kamakura, 2012).
Marketing	Marketing, customer relationship management (Ngaia *et al.*, 2009; Struhl, 2015). Customer loyalty analysis (Gans *et al.*, 2017). Finding trading rules, competition (Allen & Karjalainen, 1999).	Genetic algorithms, NLP (Allen & Karjalainen, 1999). NLP, social network mining (Gans *et al.*, 2017; Struhl, 2015).
Firm level management	Trading strategies (Tan et *al.*, 2007). Corporate strategies (Sohn *et al.*, 2003), manufacturing policies (Akhbari *et al.*, 2014). Enterprise logistics (Stock *et al.*, 2000).	Classification, prediction.

Domain	Problem treated	Method used
Supply chain optimization	Supply chain optimization (Stock *et al.*, 2000).	k-means.
Transportation management	Traffic forecasting (Zhong & Ling, 2014; Zhang *et al.*, 2013).	kNN Regression (Zhong & Ling, 2014).
Alliance-level decisions	Strategic merging decision, cross-border investments (Lau *et al.*, 2012). Finding synergies for merging and major competitors (Hoberg & Phillips, 2010).	Domain-specific sentiment analysis, business relation mining, statistical learning, evolutionary learning, business intelligence (Lau *et al.,* 2012). NLP (Hoberg & Phillips, 2010).
Knowledge transfer, innovation, knowledge management	Knowledge transfer (Li *et al.*, 2014); co-authorship networks of the US patent inventor (Balsmeier *et al.*, 2016; Choi *et al.*, 2008); knowledge management (Williams & Lee, 2009).	NLP, graph-based methods, clustering, classification, prediction, cluster analysis (Balsmeier *et al.*, 2016; Li *et al.*, 2014).

NATURAL LANGUAGE PROCESSING: TEXTUAL ANALYSIS

Management requires vast amounts of information that must be retrieved, aggregated, filtered, correlated, and analyzed from various standpoints. Many of the main sources of information come in textual form, such as corporate filings (e.g., Li, 2010a), financial disclosures (e.g., Loughran & McDonald, 2011, 2014), customer messages (e.g., Struhl, 2015; Ngai, Xiu, & Chau, 2009; Pang & Lee, 2008; Balazs & Velásquez, 2016; Mostafa, 2013; Piryani, Madhavi, & Singh, 2017; and Gans, Goldfarb, & Lederman, 2017), internal corporate documents such as corporate emails (e.g., Srivastava *et al*., 2017) and CEO diaries (e.g., Bandiera *et al.*, 2017), and patents (e.g., Hall, Jaffe, & Trajtenberg, 2001; Trajtenberg, Shiff, & Melamed, 2006; Kaplan, 2012; Li *et al.,* 2014; and Balsmeier *et al.*¸ 2016). The use of the information contained in text collections is based on methods pertaining to the domain of Natural Language Processing (NLP).

NLP is the interpretation of text and speech using automated analytical methods. Claude Shannon laid the groundwork for information theory and NLP by describing a model of communication (Shannon, 1948) and introducing statistical language models (Shannon, 1951); Alan Turing laid the foundation for artificial intelligence (Turing, 1950).

A non-exhaustive list of subfields of NLP includes language parsers and grammars, text and speech recognition, sentiment analysis (including its impacts on firm and individual behavior), document classification (including insurance fraud detection, spam detection, and news manipulation), analysis of customers' and investors' sentiment tendencies (Lugmayr, 2013), search query disambiguation (for example, handling of word associations, abbreviations and polysemy), market segmentation, customer churn modeling, and many more. Researchers in management, strategy, marketing, and accounting have all found applications of NLP relevant to understanding consumer, firm, government, and individual executive behavior.

Text Analysis Workflow

Text requires a sequence of processing stages to be quantified into variables which can then be used in regressions or classifications. A typical workflow, including the sampling and analysis step, is depicted in Figure 1. The first step of any textual analysis is to determine the sample of interest, which is generally referred to as a collection of documents, where a document refers to an observation. A document or observation can be as short as a tweet or as long as a financial report or comprehensive patent description. The computational complexities of processing text are driven by the data volume, as measured by the size of the collection of documents and the average length of a document in the collection. The analysis of documents requires the comparison of their features with those of corpuses, which are comprehensive bodies of text representing a field or a natural language.

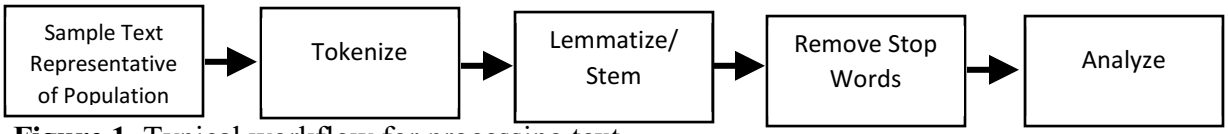

Figure 1. Typical workflow for processing text

The text preprocessing steps consist of tokenization, lemmatization or stemming, and stop words removal. Tokenization means segmenting a text, which is essentially a string of symbols including letters, spaces, punctuation marks, and numbers, into words and phrases. For example, a good tokenizer treats expressions such as "business model" as a single token, and processes hyphenation (Manning, Raghavan, & Schütze, 2008).

The other two preprocessing steps in text analysis are used depending on the purpose of the analysis. For example, when one wishes to differentiate between the specific languages used by two authors, one may wish to determine how frequently they use common words such as "the," "and," "that." These words are called "stop words" and serve grammatical purposes only. In contrast, when one is interested in sentiment analysis, words that carry semantic meaning matter; stop words are generally held not to carry semantic meaning, so for such analyses they should be removed in preprocessing. This is easily achieved with any of the standard text processing packages, which maintain dictionaries of stop words. Lemmatization, the reduction of the words to their lemma (the dictionary form of the word), helps lessen both the computational task and the duration of the analysis. Lemmatization reduces the number of words by mapping all inflections and variations of a word to the same lemma. It also disambiguates the semantic meaning of the words in a text by assigning words with the same meaning to their lemma. In sentiment analysis, for example, "improve," "improved," "improvement," and "improves" all point equally to an optimistic sentiment and share the same root; differentiating them would serve no purpose for a

sentiment analysis task. The lemmatizer does distinguish between different parts of speech and notes whether the word is used as a verb or a noun. For instance, “binding contract,” “being in a bind,” and “bind together” would resolve to distinct lemmas, although they all use forms of “bind.” A typical lemmatizer is the WordNet lemmatizer; several other stemmers and lemmatizers are described in Manning *et al.* (2008).

In other cases, information on the part of speech is not relevant for the analysis, and a simple removal of the prefixes and suffixes to reach the *stem* of the word is sufficient. The stem is the root of the word, the smallest unit of text that conveys the shared semantic meaning for the word family. For example, the stem of “teaching” is “teach.” Because stemmers do not look up meaning in the context of parts of speech, verbs and nouns resolve to the same root, which reduces complexity but at the cost of a loss of information. Stemmers are standard in any programming language or toolkit that enables text analysis. The standard stemmer (Manning & Schütze, 1999) for English language texts is the Porter Stemmer (Porter, 1980). Stemming may mask valuable information. For example, the Porter Stemmer produces on the corpus of patent titles the token “autom,” which when applied to the standard American English corpus used in the literature, the Brown corpus (Kučera & Francis, 1967), finds that the stem corresponds to “automobile,” whereas the expected word is “automate.”

While there is no generalized rule in the literature about where to use a stemmer versus a lemmatizer, all text preprocessing workflows should include at least one of the two. For complex technical texts, such as patents, lemmatization is recommended. Further background in grammars, lemmatizers, stemmers, and text processing in general can be found in the comprehensive textbook by Manning and Schütze (1999) and in Pustejovsky and Stubbs (2012).

Vector Space Model

The preprocessing steps allow us to prepare a document for consumption by a variety of numerical methods. The standard representation of a document is called the "vector space model," as each distinct word in the document becomes a feature of the document; the text can then be represented as a vector of words, with each word assigned a value. If the collection of documents is represented in an N-dimensional space, where *N* is the total number of distinct words across the collection (its vocabulary *V*), then each individual document is represented as a point within this *N* dimensional space. Each dimension (axis in the corresponding diagram) represents a different word from the vocabulary of this collection. The numerical values on the axes for each document may be calculated in different ways. There are four typical methods:

1. Binary weighting at the document level assigns a value of 1 for the word's presence in the document and 0 for the word's absence in the document. This is useful in document classification tasks, where the presence or absence of a term is what matters in assigning the document to a particular topic (Albright, Cox, & Daly, 2001).
2. *Raw Term Frequency* is the raw count of the word in the document, and does not look at the total number of words in the document or at the collection of documents. It is useful in applications of sentiment analysis, where counts of positive and negative words are taken to determine the overall sentiment of the text. The most widely used annotated dictionaries include SENTIWORDNET[17] (Baccianella, Esuli, & Sebastiani, 2010) and the University of Illinois at Chicago's Opinion Lexicon[18] (Hu & Liu, 2004).

[17] The SENTIWORDNET annotated corpus for sentiment analysis research is available at http://sentiwordnet.isti.cnr.it/. Accessed May 28th, 2017.

[18] The Opinion Lexicon consists of 6800 English words annotated with positive and negative sentiment and is freely available at the University of Chicago's website: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. Accessed May 28th, 2017.

3. The *Relative Term Frequency* (TF) is calculated as the ratio between the number of occurrences of a word in a document and the number of times the word appears in the entire collection of documents. The tokenizer preprocessing step is essential for creating the proper list of words for each document, as it removes punctuation and non-word text. Stop words could "drown" out other words that would carry more meaning for the texts under analysis, and so are removed prior to calculating TF. Inflections of a word would artificially lower the TF, which makes lemmatization/stemming critical. Importantly, the TF measure does not account for words that are common across documents.
4. Using the TF calculated at 3, one can create a separate set of weights called *Term Frequency-Inverse Document Frequency* (TF-IDF) that takes into account the number of documents in which the word appears through a separate measure called Inverse Document Frequency (IDF). Denoting the number of documents in the collection as D and the number of documents containing the i^{th} word in the alphabetically ordered vocabulary vector as D_i, the IDF is $IDF[i] = \log_2(D/D_i)$. From this definition, it is apparent that words that are common to *all* documents would lead to an IDF of 0. The TF-IDF is thus defined for each word *i* in document D_i as $TFIDF[D_i, i] = TF[D_i, i] \cdot IDF[i]$. The effect of multiplying the term frequencies for each word in each document by the inverse document frequency of that word is that words that are common across documents are weighted down, as they receive a low IDF value. However, uncommon terms that reveal specifics about a document, such as the methodological and technical terms that make a particular document unique, are weighted up by multiplication by the IDF. This is particularly useful when determining the extent of the difference between pairs of documents and is the standard method used in the NLP literature. For virtually any text analysis application that targets the unique or rare features in a document, TF-IDF is the

method of choice. For instance, patents use a highly specialized language in which common words are generally irrelevant. Younge and Kuhn (2015) performed TF-IDF on the entire patent corpus and determined the differences across patents using cosine similarity on the word vectors associated with each patent. Another application of TF-IDF in management is the comparison of corporate financial forms such as 10-Ks and 10-Qs (Li, 2010b), where words common to most firms or forms are not particularly useful for extracting features of the firm's strategy. For a comprehensive review of term-weighting methods, see Salton and Buckley (1988).

Textual Similarity Measures

The most common similarity measures used in text analysis are cosine similarity, the Pearson correlation, the Jaccard similarity, and the Dice similarity. Cosine similarity has been used to compare texts for the past 30 years (Salton & Buckley, 1988; Salton, 1991; Manning & Schütze, 1999). The cosine similarity is computed as the cosine of the angle of the pair of word vectors representing the two texts, denoted as $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$. The components of these vectors are usually word counts (Manning & Schütze, 1999, p. 301):

$$\cos(\overrightarrow{w_1}, \overrightarrow{w_2}) = \frac{\sum_i w_{1i} \cdot w_{2i}}{\sqrt{\sum_i w_{1i}^2} \cdot \sqrt{\sum_i w_{2i}^2}}$$

The cosine similarity defined above may use TF or TF-IDF as a weighting method to create the values in each vector (see Salton and Buckley (1988) for an extensive review). Unlike cosine similarity and the Pearson correlation coefficient, the Jaccard and Dice similarity indices require binary weighting for both vectors, thus acting at the set-level of the vocabularies *W* of the texts.

The Jaccard similarity measures the number of shared components between the two sets of words, and is defined (using set notation) as:

$$Jaccard(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|},$$

where W_1 , W_2 are the vocabularies for the two texts. Dice similarity is defined likewise, with the key difference that it rewards shared word pairs while simultaneously penalizing pairs of texts that share fewer pairs of words relative to the total text sizes (Manning & Schütze, 1999, p. 299):

$$Dice(W_1, W_2) = \frac{2 \cdot |W_1 \cap W_2|}{|W_1| + |W_2|}.$$

Both Jaccard and Dice indices are used in information retrieval tasks, such as classification of documents and querying (Willett, 1988). Overviews of these and other typical measures are in Manning and Schütze (1999, pp. 294-307), Salton and Buckley (1988), and Huang (2008). A survey of these measures applied to collections of short texts, such as online reviews and tweets, is found in Metzler, Dumais, and Meek (2007).

Similarity measures are key to Hoberg and Phillips's (2010) study showing that firms with products very similar in textual descriptions to those of their rivals have lower profitability, and to Younge and Kuhn's (2015) study of how patent text similarities can predict future innovation. Arts, Cassiman, and Gomez (2017) apply Jaccard similarity to the patent corpus to determine technological similarity classes and compare their classification system to the USPC patent classification system. Textual similarity measures can also be helpful in creating comparison groups and identifying new classification structures. For example, they can help find companies that create comparable products despite being in different SIC codes (Hoberg & Phillips, 2010),

companies with similar customer review sentiments, or companies that have received comparable news coverage. These comparison groups can then be used in regression analysis.

MACHINE LEARNING TOOLS AND PROGRAMMING LANGUAGES

In the category of toolkits for data mining and processing, there are several, such as WEKA, RapidMiner, and KNIME, that are convenient, due to the ease of use and fast learning curve. Several, e.g., RapidMiner and KNIME, enable users to run full machine learning algorithms applying just a drag-and-drop interface, while also providing suggestions for the best parameters for the algorithms. In RapidMiner, the recommendations are based on the inputted data and also on input from a cloud-based platform to which the software is connected, which compares the performance of various algorithms on millions of datasets. While regular programming languages that support machine learning packages, such as Python, C#, and R, provide more functionality, they are less intuitive. Data collection from the Internet is automated in some packages, a further advantage for the management researcher. In this section, I also provide an overview of two natural language processing packages useful to management researchers working with text: NLTK and AYLIEN, a cloud-based toolkit. Under languages supporting machine learning, I briefly survey Python, C#, and Java. A review of these and other languages and tools is available in Louridas and Ebert (2017).

A programming task in these tools may reduce to selecting a sequence of prebuilt operators to create a sequence of linked operators that forms a process. Figure 2 depicts a process that computes the Term Frequency vectors for a collection of text documents in RapidMiner. The input for the collection of documents is specified by an operator from the Data Access list. The Select Attributes operator allows selections for the columns to be used as input variables to the text processing algorithm. The actual document processing occurs in the Process Documents operator, which can

take a wide variety of inputs, for example from a collection of files, from Twitter, or from a custom website.

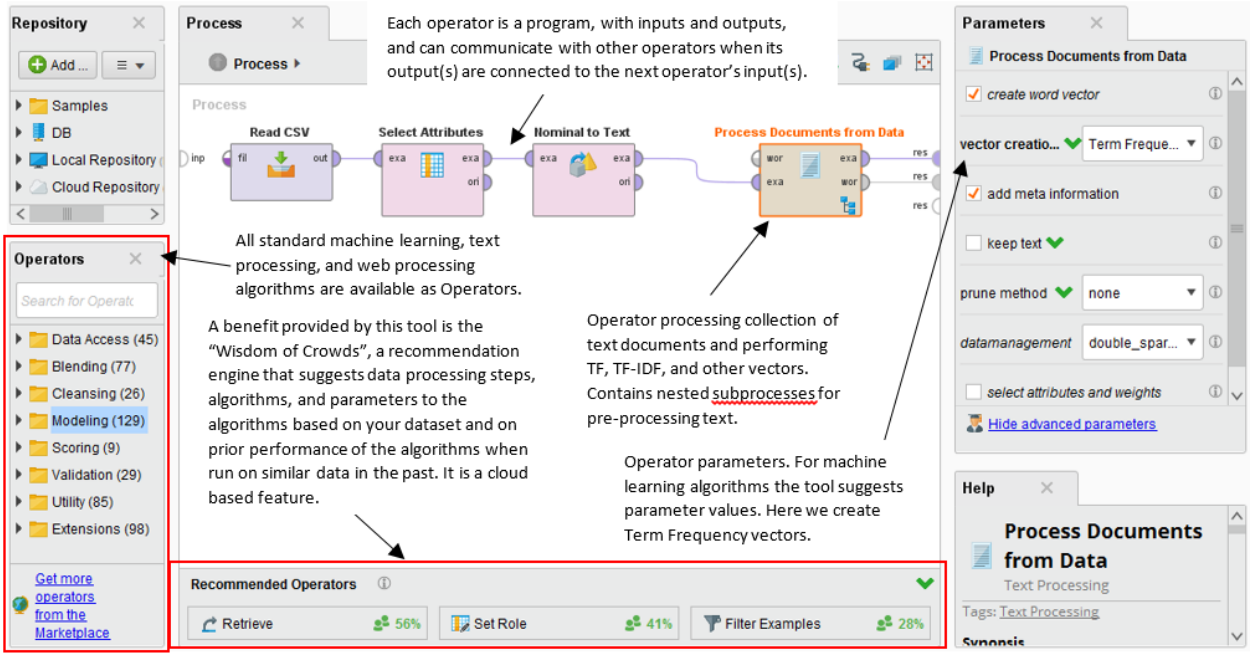

Figure 2. Example of a typical workflow for processing a collection of text documents (with overview of the RapidMiner interface).

Most toolkits include naïve-Bayes, tree-based algorithms, nearest neighbor, and support vector machine algorithms (discussed in the general-purpose machine learning section of the chapter), neural networks, and others. Toolkits also provide standard statistical models and methods, including a suite of regression, segmentation, and correlation operators. Toolkits offer a wide variety of web mining tools (Kotu & Deshpande, 2014), including tools that gather data from any website given search parameters, gather data from websites with authentication, gather data from Twitter, and collect emails from an email server. The latter two have proven especially useful data sources for recent strategy research. For instance, Gans *et al.* (2017) analyzed sentiment in customer tweets to predict firm behavior. Srivastava *et al*. (2017) applied a tree-based machine

learning approach to a firm's email server and applied a tree-based approach to determine how well employees matched the firm's email culture, and how differences in culture may impact employee turnover. The methods in these two papers could be implemented in current toolkits such as RapidMiner with just a few dragged-and-dropped operators, without the need to learn a programming language.

A unique feature of some toolkits compared to programming languages with support for machine learning is the ability to incorporate into the algorithms previous successful experience and knowledge from other sources. The use of the "wisdom of crowds" has been applied in many fields, such as biology, medicine, and NLP (Savage, 2012). A "wisdom of crowds" cloud engine (see the lower part of Figure 2, RapidMiner implementation) is a useful complement; it provides suggestions for parameter values for the operators as well as a sequence of operators that construct a program to analyze the inputted data.

The ability to visualize data and results is built into many tools, such as Tableau, Qlik, SAS, MATLAB, and RapidMiner. However, RapidMiner is more limited in its visualization capabilities than visualization tools such as Tableau and Qlik or visualization packages such as D3 or Python's Matplotlib. Good overviews of MATLAB for finance and economics include Anderson (2004) and Brandimarte (2006).

Statistical languages such as R provide machine learning packages, but their implementation time is not as fast as that of toolkits. R requires individual packages for different algorithms, as each package is relatively limited in scope (packages are available at CRAN). For example, "rpart" is used for basic classification algorithms, but ensemble methods require additional packages, like "party" or "randomforest." Other packages are built around specific algorithms, such as neural networks in "nnet" and kernel-based machine learning models in "kernLab." Generally, these

require a bit more research and learning than the prebuilt packages in MATLAB, RapidMiner, or SAS. Two good resources for working with machine learning algorithms in R are Friedman, Jastie, and Tibshirani (2001) and the associated datasets and packages and the UC Irvine Machine Learning Repository (Lichman, 2013).

SAS makes possible the statistical data analysis, data management, and visualization that are widely used in business intelligence. It claims a more accessible interface than R, with targeted packages for specific fields. Such specialized packages are not free but provide a wide array of tools, as in the case of Enterprise Miner, which provides a comprehensive set of machine learning tools, overviewed in Hall *et al.* (2014), the closest equivalent in terms of functionality to the tools already discussed. Like RapidMiner and the freeware R, SAS has a free academic edition. The general-purpose programming languages Python, C#, and Java all have a variety of machine learning, text analysis, and web mining packages. For example, in Python, the typical packages covering machine learning functionality include NLTK for natural language processing, scikit-learn and pylearn2 for machine learning methods, beautifulsoup for web parsing, pandas for data parsing from files, and Matplotlib (MATLAB-like interface) and Seaborn for data visualization. For C#, a good library for machine learning is Accord.NET, and a good library for natural language processing is Stanford's CoreNLP. Machine learning package examples for Java include the user-friendly freeware Weka and Java-ML.

In terms of packages specifically targeted to natural language processing, NLTK is a comprehensive text analysis platform for Python, whereas AYLIEN is a cross-language cloud based text processing toolkit with advanced sentiment analysis, news parsing, and named entity extraction abilities. NLTK is better for corpus analytics, as it incorporates over 100 text corpora

from different fields,[19] contains a lemmatizer based on WordNet, and has extensive functionality for sentence parsing based on grammars. For an exhaustive overview of NLTK capabilities and examples, see Bird, Klein, and Loper (2009). NLTK is used in the corpora and Zipf's law section of this chapter.

For the management researcher interested in easily collecting data about firms and then analyzing the data for sentiment or for entity extraction (locations, individuals, company names, product names, currency amounts, emails, or telephone numbers) from news sites, Twitter, documents, or websites in general, AYLIEN is available as a text extension for RapidMiner and as a Python, Java, and C# package. The news and Twitter parsers allow the user to connect these entities to collections of text documents, which can then be linked to events like stock prices or product launches and assigned a sentiment value through the prebuilt sentiment analyzer.

Sentiment Analysis and the Naïve-Bayes Classifier Using NLP

Investor sentiment is known to affect stock returns (Lee, Shleifer, & Thaler, 1991), and investors themselves are known to be influenced by the sentiment of news articles (Tetlock, 2007; Devitt & Ahmad, 2007), by the sentiment of conventional media (Yu, Duan, & Cao, 2013), by social media (Bollen, Mao, & Zeng, 2010), and by nuances of optimism about future events as reported in standard financial filings (Li, 2010b). Attitudes and sentiments are detected by counting "positive" and "negative" words and expressions, using specific "bags" (sets) of sentiment/opinion words in lexicon-based detection methods (such as in Taboada *et al*. (2011) or Ravi and Ravi (2015)), and calculating sentiment scores as the ratios of these counts (Struhl, 2015). The second class of methods for sentiment detection pertains to machine learning. Various types of supervised

[19] For a list of current linguistic corpora included with NLTK, see http://www.nltk.org/nltk_data/. Accessed May 29th 2018.

classifiers are used in the literature to mine for the sentiments in a text, such as neural networks (NN), support vector machines (SVM), rule-based (RB) systems, naïve-Bayes (NB), maximum entropy (ME), and hybrids. Ravi and Ravi (2015) and Tsytsarau and Palpanas (2012) provide details on the classifiers and related machine learning techniques in opinion mining. N-grams, which are uninterrupted sequences of N tokens, are often used in sentiment analysis to classify the sentiment of expressions. In the cases of online data sources, tokens may include punctuation constructs in the form of emoticons, and n-gram analysis takes into account the affect that an emoticon carries, such as through the use of bi-grams (pairs of tokens) to analyze consumer behavior and sentiment with regards to actions in the airline industry (Gans *et al.*, 2017).

In most sentiment analysis applications, a classification decision must be made regarding the type of sentiment (positive, negative, or neutral) at the document level or the sentence level. The typical classifier used in this context (Li, 2010b; Gans *et al.*, 2017) is the naïve-Bayes, a fast, general-purpose classifier popular in sentiment analysis (e.g., Pang & Lee, 2004, 2008; Melville, Gryc, & Lawrence, 2009; Dinu & Iuga, 2012).

The naïve-Bayes classifier works by using Bayes' rule for each classification decision under the assumption that *all predictors* are independent of each other. The name of the classifier is drawn from this assumption, which yields a certain naiveté in many situations but also makes this the simplest classifier, with no parameters to tune. The lack of parameter tuning makes it one of the fastest classifiers, which is especially useful in problems in which real-time analysis is needed, such as stock trading and question-answering bots. The naïve-Bayes algorithm calculates the posterior probability for each class given a predictor and picks the class with the highest posterior probability as the outcome of the classification.

Corpora and Zipf's Law Usefulness in Patent Sub-Corpora Comparisons

Business applications such as marketing sentiment and shareholder analysis require large corpora composed of collections of messages and documents. A linguistic corpus is a "systematically collected" set of "machine-readable texts" "representative of 'standard' varieties of [a language]" (Leech, 1991, p. 10; Pustejovsky & Stubbs, 2012, p. 8). A corpus should be a representative sample of the overall language, which may be a natural language or a specialized language like those used in patents, individual scientific fields, financial reports, consumer reviews, or short online texts (Tweets, product descriptions, firm mission statements). A comprehensive list of the most used and freely available corpora is provided in the appendix of Pustejovski and Stubbs (2012); the library NLTK, discussed in the prior section, also provides a growing set of free linguistic corpora. It is widely accepted that the standard American English corpus is the Brown corpus (Kučera & Francis, 1967), which was created as a representative sample of the English language by a group at Brown University in the 1960s. The Brown corpus features balanced coverage of the different genres of the time (Manning & Schütze, 1999, p. 19), and covers about 1 million words. The optimal corpus for a given field is a corpus formed by collecting a complete population of the texts that define the field.

The selection of an appropriate corpus for the research setting is essential, as using a general-purpose corpus can lead to misleading results (Li, 2010a). Corpora may be tagged and annotated to enhance their analytical usefulness. For example, words may be tagged with their part of speech to enable statistical analysis of the inherent grammar of the language, as in the case of Treebanks, whose gold standard is the Penn TreeBank (Marcus, Marcinkiewicz, & Santorini, 1993). Such tags can then be used as an input to allow a classification algorithm to learn to classify a wider body of text, as is the case of the Reuters corpus, which collected a balanced sample of news articles

spanning 90 topics and classified each article (for an overview of Reuters corpus uses, see Sebastiani, 2002). For a methodological overview of the composition of a linguistic corpus, see Pustejovsky and Stubbs (2012), and for a programmatic reference for working with corpora, see Bird *et al.* (2009) and Marcus *et al.* (1993).

Zipf's law (Zipf, 1949) is a descriptive instrument and a benchmark in the analysis of large text collections. Zipf's law states that if one were to order the words of a natural language by their frequency of appearance in the language and name the ordered position of each word as its *rank*, the relationship between the frequency *f* and the rank *r* would be (Manning & Schütze, 1999, p. 24):

$$f \propto \frac{1}{r^{\alpha}}, \alpha > 1.$$

In a log-log plot of rank and frequency (a standard visualization for corpora), Zipf's law would yield a line with a slope of about -1. The lower end of the distribution is dominated by the most infrequent words, which are helpful for defining a text for searches and in similarity analysis. Zipf's law is just an approximation of a language; the Brown corpus, for example, deviates at the tails of the distribution from the ideal (see Figure 3). Despite this limitation, the log-log frequency-rank construction remains helpful for comparing the distributions of linguistic corpora and for identifying differences between a natural corpus such as Brown and corpora that are field-specific. Such differences matter for accurately understanding sentiment in financial disclosures in the form of forward-looking statements, which do not follow typical sentiment dictionaries and require training a classifier on a specialized collection of documents (Li, 2010b). Similarly, in the case of the language of online reviews, a specialized corpus must be created to properly classify sentiment (Gans *et al*., 2017). In the patent corpus, specialized technical language yields a very different

distribution from that of a natural language corpus such as Brown. Figure 3 shows the comparison between the standard Brown corpus, the US-granted patent abstracts corpus, the US-granted patent titles corpus, and the US-granted patent claims corpus (all patents issued between 2007 and November 2017). The heads (due to stop words) and the tails of the distributions differ in the Brown and the three patent corpora. For the titles corpus, for example, relatively technical words such as "method," "device," "system," "apparatus," and "control" appear at the top of the distribution. This suggests that for patent similarity analysis, the most frequent (and thus least informative) words are different from those in a regular English language corpus, and that the distributions of the corpora should be determined in advance. Zipf's law is a regularity that also holds for city population sizes (Gabaix, 1999; Glaeser *et al.*, 1992), firm size by employee number (Axtell, 2001), and firm bankruptcies (Fujiwara, 2004).

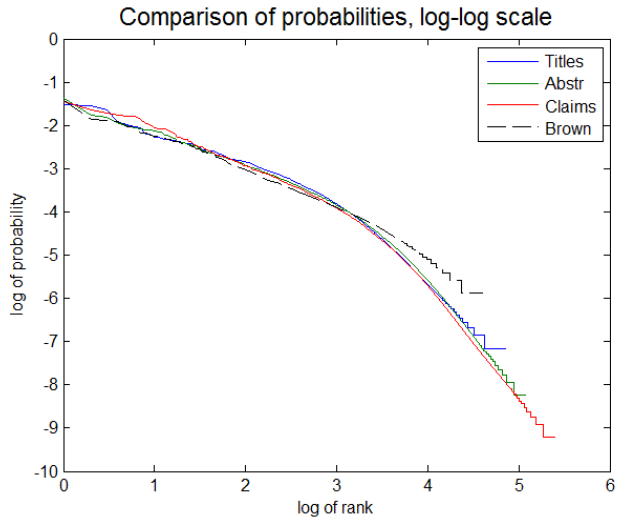

Figure 3. Log-Log representation of the Brown Corpus exhibiting Zipf's law, as compared to the patent titles corpus, the patent abstracts corpus, and the patent claims corpus.

Topic Modeling – Applications to Patent Texts

Topic Modeling encompasses a set of methods that extract themes from bodies of texts; these methods can be used to classify texts as well as to summarize texts. Topic models are conceptually a layer above the vector space model described earlier, as topic models also look at the relationships between words, such as words often occurring in close proximity, whereas the vector space model looks at raw term counts. Probabilistic topic modeling approaches assume that words follow a distribution over topics and that topics follow a distribution over documents in a corpus. In other words, documents are obtained through a *generative process* in which topics are generated from a distribution and those topics themselves are generated from another distribution of words. The approach that has become a standard in the field is the Latent Dirichlet Allocation of Blei *et al.* (2003), in which the generation of every *document d* in a corpus is described as a probabilistic generative process (simplified below; for the full derivation, see Blei *et al*., 2003, pp. 996-1006):

1. For every document, draw number of words of document *d* from a Poisson distribution
2. For every document, draw the proportions of the topics for document *d* from a Dirichlet distribution
3. For every word in document *d*:
    a. Draw the topic the word is assigned to (from a multinomial)
    b. Draw the word itself (from the multinomial of the topic)

The assumption here is that the number of topics of the corpus (the entire collection of documents) is known. The documents are the observations; the topics, topic-document, and word-topic assignments are all hidden variables. There are multiple convergence approaches, of which one of the most popular is collapsed Gibbs sampling (Porteous *et al.*, 2008; Asuncion *et al.*, 2009; Xiao *et al.*, 2010). Variational Bayes has also been growing in popularity due to its speed on very large

online corpora (Hoffman *et al.*, 2010). The implementation of LDA using variational Bayes is used in the application on the patent corpus in this subsection from the Gensim[20] Python package.

For the application in this section, I have chosen to run a topic model on the claims corpus of the US Green Technology Pilot Program Patents (a program that ran between 2009-2012 at the US Patent and Trademark Office, granting accelerated examination to patent applications on green technologies, and analyzed further in Chapter 3), as well as on the claims corpus of all granted patents between 2009 and 2012. The purpose was to draw a comparison between the two corpora. My implementation was in Python and imported the libraries NLTK for Natural Language Processing tools and Gensim for topic modeling. The output in Table 2 shows that even with a limited number of words per topic (set as five per topic for this example), as expected, there are clear differences in the topics of the documents (the topics are ordered based on their rank; the table shows the top ten most represented topics in the corpora):

Table 2. Top topics found through application of LDA topic modeling to the claims corpus of green technology pilot program patents (left column) and to the claims corpus of all 2009-2012 patents. Topics were limited to five words per topic.

Topic Rank	LDA Green Technology Patents Top Topics	LDA All Patents 2009-2012 Top Topics
1	(wind, blade, turbine, fuel, surface)	(data, end, structure, group, level)
2	(gas, side, solar, current, fluid)	(image, data, set, signal, configured)
3	(light, wind, turbine, gas, flow)	(data, memory, circuit, configured, element)
4	(gas, power, wind, heat, turbine)	(data, group, control, signal, member)
5	(power, signal, canceled, cell, control)	(apparatus, member, end, image, configured)
6	(configured, wind, turbine, support, fluid)	(layer, control, group, light, apparatus)
7	(wind, turbine, fuel, configured, air)	(layer, data, user, configured, image)
8	(heat, liquid, stream, gas, wind)	(power, side, control, group, signal)
9	(voltage, energy, engine, fuel, light)	(body, material, value, apparatus, end)
10	(material, substrate, surface, layer, fluid)	(acid, surface, signal, layer, material)

[20] Gensim is a topic modeling toolkit available for the Python programming language available at https://radimrehurek.com/gensim/. Accessed April 10th 2018.

A nonparametric extension of LDA is HDP – Hierarchical Document Process – which solves the limitation of LDA, in which the number of topics had to be prespecified (Teh *et al.*, 2005). HDP can be utilized to determine the optimal number of topics for a given corpus. In our particular data, the topics obtained through HDP appear to be more interpretable. HDP was run in comparison with LDA for both the Green Technology Pilot Program patent claims corpus (Table 3) and the regular 2009-2012 patent claims corpus (Table 4). However, the reader should not draw general assumptions about the interpretability of topics from this limited example; the interpretability of topics generated by topic modeling remains an active area of research in computer science, with approaches shifting to neural networks such as Long-Short Term Memory Recurrent Neural Networks algorithms (Ghosh *et al.*, 2016; Li *et al.*, 2016). The results in Tables 3 and 4 utilize the variation of HDP from Wang *et al.* (2011), implemented in Python's Gensim library.

Table 3. Comparison of top topics identified through LDA versus top topics identified through HDP for the Green Technology Pilot Program patent claims corpus. Note the HDP topics provide a more granular level of detail and appear more interpretable.

Topic Rank	LDA Green Technology Patents Top Topics	HDP Green Technology Top Topics
1	(wind, blade, turbine, fuel, surface)	(turbine, flow, power, gas, master)
2	(gas, side, solar, current, fluid)	(power, energy, output, voltage, input)
3	(light, wind, turbine, gas, flow)	(signal, rotor, power, blade, current)
4	(gas, power, wind, heat, turbine)	(wind, power, turbine, layer, defrost)
5	(power, signal, canceled, cell, control)	(metal, molecular, sieve, zeolitic, catalyst)
6	(configured, wind, turbine, support, fluid)	(wind, power, turbine, output, direction)
7	(wind, turbine, fuel, configured, air)	(power, configured, signal, converter, input)
8	(heat, liquid, stream, gas, wind)	(wind, power, generation, step, radius)
9	(voltage, energy, engine, fuel, light)	(element, coupling, region, lamp, end)
10	(material, substrate, surface, layer, fluid)	(power, deflector, stator, gas, less)

Table 4. Comparison of top topics identified through LDA versus top topics identified through HDP for the 2009-2012 patent claims corpus.

Topic Rank	LDA All Patents 2009-2012 Top Topics	HDP All Patents 2009-2012 Top Topics
1	(data, end, structure, group, level)	(apparatus, image, member, data, surface)
2	(image, data, set, signal, configured)	(data, signal, image, memory, display)
3	(data, memory, circuit, configured, element)	(signal, data, layer, user, based)
4	(data, group, control, signal, member)	(control, memory, cell, signal, acid)
5	(apparatus, member, end, image, configured)	(data, image, computer, based, apparatus)
6	(layer, control, group, light, apparatus)	(data, light, surface, apparatus, layer)
7	(layer, data, user, configured, image)	(data, configured, computer, network, user)
8	(power, side, control, group, signal)	(end, power, member, configured, surface)
9	(body, material, value, apparatus, end)	(layer, surface, material, region, element)
10	(acid, surface, signal, layer, material)	(data, signal, user, value, network)

Topic models on large bodies of texts, such as patents, firm press releases, or even firm internal documents, are useful for performing a finer comparison of knowledge transfers, competition, and differences in firm strategies. This is an area of active recent research in strategic management (Younge & Kuhn, 2016; Arts *et al.*, 2017), and there is an opportunity to create a similarity measure and comprehensive dataset. A natural extension of this section is to utilize the topics generated by a topic model on patents to generate topic vectors for each document in a collection and calculate the cosine similarity on such vectors rather than on words. Work is underway on a revised version of this chapter as an HBS working paper that will include a topic model-based similarity calculation for the patent corpus.

GENERAL-PURPOSE MACHINE LEARNING METHODS AND APPLICATIONS

Learning Types: Taught Versus Self-Educated

The "learning" in machine learning is that of an algorithm that tweaks the parameters of a model of data to reach a goal based on an optimization criterion. The goal varies according to the application—it could be winning a game of chess, predicting customer purchasing behavior with

an accuracy above a set requirement, segmenting the data until a mathematical minimization criterion is achieved, or reaching a certain population composition after evolving over thousands of generations. Learning falls into four main categories: unsupervised learning, supervised learning, reinforcement learning, and evolutionary learning, with the last two often included in supervised learning. Machine learning uses a model that ties the outcome variable (which may be referred to as the *target*) to the explanatory variables (which are referred to as *parameters* in the model). In the case of *supervised learning*, the algorithm has knowledge of the values of the explanatory variables that lead to a given outcome in the existing data, or a "teacher" gives input that corrects erroneous decisions taken by the algorithm. For instance, in the case of a loan decision, a bank employee may override an algorithm decision. Direct human intervention in the form of teaching is frequently seen in artificial intelligence applications (image classification with the subset field of handwriting recognition, which is now ubiquitous in the postal and banking sectors; speech recognition; spam filtering) but is less relevant for management applications. In the case of social science research, one might have a small subset of the data manually coded (outcome variable known) and need the algorithm to code the remainder of the data (outcome unknown), or the data may already be coded in terms of the correct outcome variable values, yet the mechanism producing these outcomes may be unknown.

In the case of unsupervised learning, none of the data are tagged and there is no human intervention to help the algorithm adjust along the way. Unsupervised learning covers data segmentation problems familiar in statistics such as clustering (k-means is a typical method here) or principal component analysis (PCA). Clustering is popular in marketing, particularly in consumer market segmentation questions (as in Punj & Stewart, 1983; Schaffer & Green, 1998; Wedel & Kamakura,

2012), and an approach similar to PCA has been used recently to determine components of CEO behavior (Bandiera *et al*., 2017).

Reinforcement learning rewards the machine if it reaches the correct outcome, but does not reward individual steps. It is a form of machine learning for multistage games, in which no one move necessarily leads to the desired outcome, but there exist multiple winning paths. The machine becomes better after each completed game.

Evolutionary learning is inspired by biological mutation, selection, and reproduction in populations. The drawbacks of evolutionary algorithms include the high mathematical complexity that makes computation time a factor to consider and the fact that the solutions may not necessarily be as intuitive or easily interpretable as those of other methods. Their applications include financial trading (Allen & Karjalainen, 1999).

K-means: An Example of Unsupervised Learning

K-means is a typical case of unsupervised learning and thus of knowledge discovery. The method requires as an input a good guess of the number of classes, k, which the researcher should make before running the algorithm. In this respect, it is a method of partial knowledge discovery with partial supervision, as the number of clusters is "taught" to the machine. It is similar in principle to k-Nearest-Neighbors (also summarized in this chapter) as it uses distances, yet it is based on cluster centers called centroids. Initially, the k clusters consist of a single element each; these elements are picked randomly from the data (if no good guess can be made about how to pick them) in the initialization step. At this stage, the selected data points become the "centers" of the newly formed clusters. In the next step of the most elementary form of the procedure, a new data point is randomly selected, the distances between the newly selected data point and the centers of

the clusters are determined, and the new element is assigned to the closest cluster. The cluster now contains two data points, and its centroid is computed as the average coordinate values of the two members of the cluster. Subsequently, new data are selected randomly and the previous steps are repeated. Alternatively and more frequently, in the second step all data points are assigned to one of the clusters and contribute to the computation of the new centroids. During the run of the algorithm, the centers of the clusters continuously migrate, eventually tending toward fixed positions. When the centers of the clusters stabilize, the machine has learned the statistics of the data and the process is stopped. The k-means method is relatively simple algorithmically, but time-consuming.

The uses of k-means in the management literature have primarily been in the preparatory stages of analysis. In their study of the relationships between knowledge management strategies and organizational performance, Choi, Poon, and Davis (2008) used k-means to determine clusters of companies as a first step. Another example is Ngai *et al.* (2009), who discussed k-means in the context of customer clustering, while Chiu *et al.* (2009) used the method for market segmentation. Wang (2009) performed a detailed analysis of various clustering methods in market segmentation based on published research and concluded that k-means is not the most robust technique, although it behaves reasonably. Wang suggested the use of hybrid kernel-based methods for customer relationship management applications, especially when the target clusters are overlapped and outliers are present.

Principal Component Analysis: Unsupervised

Principal component analysis (PCA) is a method used in multivariate data analysis to sort out the input variables that play the most important role in explaining the variance of the results. It is used primarily when there are numerous input variables with varied degrees of importance in explaining

a process, and when one suspects that some of these variables play no or little role. The technique is commonly used *before* machine learning methods are applied to reduce the dimensionality of the problem and accelerate computations by projecting a massive multidimensional space into a subspace of much lower dimension. The applications of PCA in management are numerous. For example, Van de Vrande *et al.* (2009) used PCA "to reduce the number of dimensions in data and applied cluster analytic techniques to find homogeneous groups of enterprises" (p. 430). Elenkov, Judge, and Wright (2005) recalled two other applications of PCA, one in providing "support to the typology of product-market and administrative innovations" (p. 666) and the second in developing "measures for Product-Market Innovation and Administrative Innovation" (p. 673). Lamberg *et al*. (2009) "employed PCA to illustrate the movement of […] organizations in the competitive landscape" (p. 55). PCA has also been used to analyze relationships between corporate vision statements and employee satisfaction (Slack *et al*., 2010), extract factors from survey questions in investment exit decisions (Elfenbein *et al*., 2016), analyze board decisions regarding CEO compensation (Zhu, 2013), and create board independence measures (Lange *et al.*, 2014). Clustering and regression studies that work with large numbers of variables may need to apply PCA before the core analysis is performed.

Training, Testing, Cross-validation in Supervised Learning: Core Concepts

A supervised algorithm requires a pre-tagged dataset in which the correct outcome for each data point has already been made available. This pre-tagged dataset is called the training set, as it is used by the algorithm to learn and update its parameters. The second dataset is called the test set, and is used for validating the model determined during the training portion. Running a learning algorithm once does not eliminate the possibility that the model may avoid generalizing to new data and instead merely overfit the training data. To solve this problem, the data are randomly

sampled into different subsets and the model is run multiple times on different splits of the data, a method called cross-validation. Cross-validation involves measuring algorithm accuracy and comparing different runs of the algorithm across the various splits of the data to optimize model parameters. Multiple flavors of cross-validation are standard in any machine learning toolkit, including holdout cross validation, k-fold cross-validation, and leave-some-out multi-fold cross-validation.

The k-fold cross-validation approach is the typical one in most applications (Manning & Schütze, 1999). In this method, the data are randomly partitioned into k mutually-exclusive subsets and the algorithm is run k times, with each run on a different set of k-1 subsets joined as a training set and with testing done on the remaining subset. The k runs thus produce k different parameter sets for the algorithm, and the classification performances of these runs can be compared to each other. K is normally selected to be at least ten folds. Out of the k runs, the best-performing algorithm is selected as the outcome of the cross-validation process. The performance of a machine learning algorithm, however, involves more than just accuracy. Measuring it requires a few more concepts, which I discuss in the following section.

Classification and Accuracy Measures

The simplest measure is *classification accuracy*, defined as the sum of the true positives and true negatives divided by the total number of classification decisions (sum of the true positives, true negatives, false positives, and false negatives.) A frequently used measure is *precision*, which is defined as the true positives divided by the sum of the true positives and false positives. *Recall* is defined as the true positives divided by the sum of the true positives and false negatives. The *Confusion Matrix* is a simple 2 by 2 matrix representation of the four standard accuracy metrics by which to measure the performance of a classifier: True Positive Rate, False Positive Rate, False

Negative Rate, and True Negative Rate. It can be used to compare different algorithms on the same data and tradeoffs of the various algorithms.

The typical method used to compare different runs of the same classifier (e.g., in cross-validation) or different classifiers run on the same data is the Receiver Operating Characteristic (ROC) curve, a plot of the true positive rate versus the true negative rate. It is also applied to compare the performance of classifiers against chance, which in an ROC plot is the diagonal. The area under the ROC curve is one of the most frequently encountered measures of classifier performance, and is called the *Area Under the Curve* (AUC). The best possible model would have an AUC of 1, as it would be a perfect detector of true positives and return no false positives.

K-Nearest-Neighbors: A Simple Supervised Learning Method

The k-Nearest-Neighbors (kNN) algorithm is one of the simplest of the supervised learning methods. The closest k training data are chosen, and the majority vote across these k wins and is assigned as the classification for the new data point. The choice of k matters, as values that are too small, such as k=1, overfit, whereas large values of k (k=21, for example) take considerable computational time. Using cross-validation, one can easily find the value of k with the lowest misclassification error. Unlike the simplest classifier, naïve-Bayes, the kNN does not assume independence between the predictors and can take on any decision boundary shape (the decision boundary separates distinct classes in the predictor space). If speed of prediction as measured by time to classify new data matters for the application, kNN is a worse choice than naïve-Bayes, as it is slow at prediction, which matters in financial market applications and other applications where near-real-time responses are needed. Although kNN is biased, corrections can be applied (Magnussen *et al.*, 2010).

Decision Trees: Supervised Learning

A tree is a graph with a node designated as a *root*, any two nodes of which are connected by only one path. All nodes except the terminal nodes, called *leaves,* branch out. Each branching represents a decision. A typical type of decision tree is the binary tree, where each decision yields exactly two choices. Trees with more than two branches per node are used as well, though this discussion centers on the binary classification tree. The decision tree can be used to classify any kind of categorical outcome. Key benefits of decision trees compared to other methods are that they are intuitive and can be used to generate an *induction rule set*, the set of mutually exclusive classification rules that yield every possible outcome in the tree.

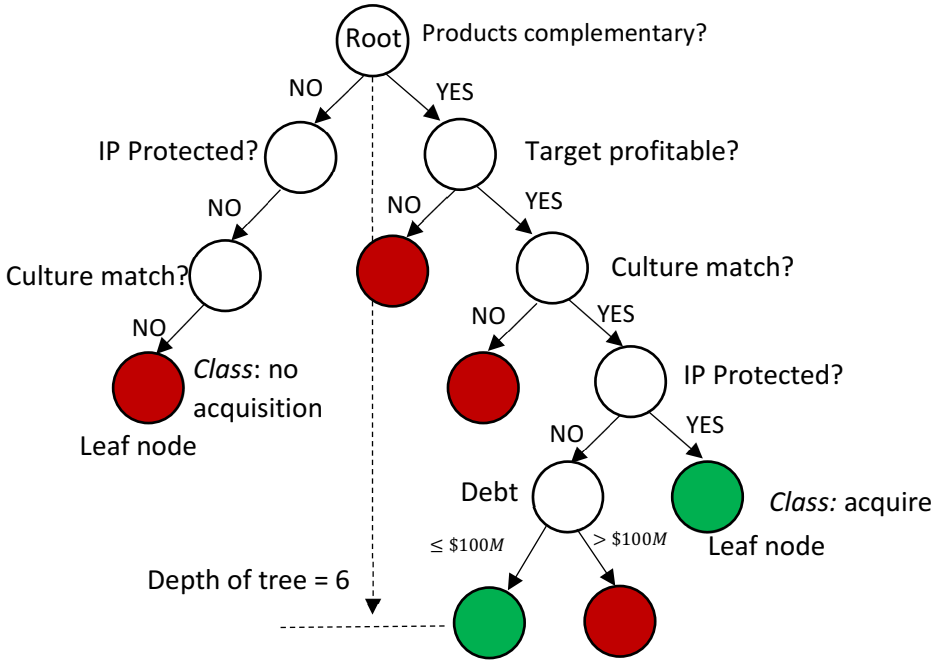

Figure 4. An example of a decision tree with a variety of path lengths

In Figure 4, I exemplify a few decision paths for an organization looking at an acquisition. This is a simplified example of the decision paths a company might take in acquiring another: some of the decisions might involve binary variables (e.g., is the target company profitable or not?) or numerical variables (e.g., debt load). Such a combination poses no difficulties for a decision tree model, which splits the feature space of the data based on the individual variables.

Each path through the tree leads to a *leaf node* that allots all data satisfying that decision path to a particular class. In the crude example from Figure 4, there are two decisions, to acquire or not acquire, and a set of variables, some categorical and one numerical, that have different orders of importance depending on the path. Leaf nodes *can occur at any depth in the tree*, as some decision paths may be longer than others. On each path, the nodes closer to the root are of greater importance. If the leaf nodes are *pure*, the elements found in each of these leaf nodes are homogenous in their characteristics. To understand a standard decision tree algorithm, it is helpful to introduce the notion of *entropy* as the minimum amount of information necessary to transmit all possible outcomes in a random variable X (in other words, the minimum size of a binary message):

$$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x)$$

The entropy for a binary classification with two classes x_1 and x_2 is

$$H(X) = -p(x_1) \cdot \log_2 p(x_1) - p(x_2) \cdot \log_2 p(x_2)$$

with a maximum entropy for $p(x_1) = p(x_2) = 0.5$ corresponding to a random toss. The amount of entropy increases with the number of possible classes. The splitting of nodes is decided by the maximal reduction in the entropy, i.e., for which the largest information gain (IG) is achieved:

$$IG(X, A) = H(X) - H(X, A).$$

For a complete derivation, see Chapter 16 of Manning and Schütze (1999, pp. 575-597).

The process of node splitting is iterative: at each step the algorithm decides whether to split on a new attribute based on whether the entropy post-split is smaller than the entropy pre-split. This is the principle of the simplest decision tree algorithm, the ID3 iterative algorithm (Quinlan, 1979). There are numerous variations of decision tree algorithms (Marsland, 2015).

The depth of the tree is learned from the data. In the absence of a stop criterion, however, the tree could generate enough splits to perfectly model the input test data, thus resulting in an overfit model. As described earlier, cross-validation is a typical approach to prevent overfitting (Elith, Leathwick, & Hastie, 2008).

Decision trees have been used for survival analysis as an alternative to logistic regression; a typical example is the Titanic survival data. For a comprehensive overview of this example and a comparison of logistic regression to decision trees, see Varian (2014). Recent work has also shown that decision trees can be used as an alternative to propensity score matching, as in Westereich, Lessler, and Funk (2010).

Decision trees and their resulting if-then *rule sets* can help both to understand processes in the data and to design new studies. For instance, past legal decisions may be modeled using decision trees, and the results may be applied to predict how firms may respond to a legal decision or use litigation as a tool in competitive behavior. The newly-published PACER patent litigation data set may be amenable to such analysis (Marco, Tesfayeus, & Toole, 2017). Decision trees can be employed to classify decisions in corporate documents if combined with NLP tools to extract variables from emails, memos, and financial filings. The lessons learned from such analyses may help

organizational behaviorists design surveys based on features and decision paths extracted from data.

Forests, Bagging, Boosting: Supervised Learning

Some machine learning models are dependent on changes in the initial conditions in their data or their input parameters. Ensemble learning methods aggregate many runs of these models to generate a more generalizable model. One of the simplest such approaches is the *bagging* method, which combines bootstrapping with aggregation. Essentially, many different runs using different training data for each run through bootstrapping are aggregated through a majority voting system (or other criterion) to generate a new model based on these aggregated parameters. *Boosting* assigns higher weights to misclassified data, such that subsequent runs of the algorithm sample more of the misclassified data points and thus focus on reducing these misclassifications (for an approach using boosted regression trees, see Elith *et al.*, 2008). The different runs are aggregated through voting. Bagging and boosting are general-purpose ensemble methods. The *random forest* technique, by contrast, focuses solely on decision trees, generating a number of pre-specified decision tree models, each with a randomly selected number of attributes from the data. The number of attributes chosen is less than the dimensionality of the data. Each decision tree receives an identically sized but randomly sampled training set. Finally, for each data point, classification is determined as the majority vote of all the trees' decisions for that data point. This ensemble method often outperforms simple cross-validation for decision trees. Random forests do this, and are also highly resilient to outliers and noise in the data. An in-depth comparative discussion of these method and their variants applied to credit scoring is in Wang *et al.* (2012).

Application Issues and Examples

An example additional to the applications of algorithms using patent data is available in the appendix to this chapter, in which a text analysis of firm mottos is used to determine clusters of competitors. Sentiment analysis can be used also for analysis of firm press releases, as well as other firm documents such as internal memos, and can prove a powerful tool for the strategic management researcher.

DISCUSSION AND CONCLUSIONS

Methods pertaining to natural language processing, decision trees, clustering and classification have become necessary instruments for strategy and management in domains such as multinational corporations, international commerce, financial markets, alliances and mergers, corporate governance, supply chain optimization, transport management, banking, and knowledge transfers. This chapter surveyed part of the field of machine learning for methods relevant to recently accepted practices in strategy and more broadly in management research. Next, the chapter focused on clustering, classification and decision methods and their relation to prediction and other decision tools as used in strategy and management. It listed the theoretical and applied limits of the methods in strategic management and mentioned current and future research directions.

To maintain a competitive advantage, management teams need to understand the tools and methods of machine learning suitable for generating from the raw data information about the categories of customers and competitors, the patterns of their behavior in the market, their relationship with and sentiments toward the firms, and the trends of the above. The competitive edge of a firm may have its roots in the amount of data it can obtain and process, in the quality and depth of the data processing, and in the ability of the management teams and strategists to ask

questions of and cooperate with the data analysts. In turn, a fruitful collaboration between them requires an understanding by the strategists of the methods and tools in machine learning.

This chapter summarized some of the tools in machine learning, particularly in using natural language processing methods, decision trees, clustering, and classification methods. Further, the chapter aimed to show how to combine language processing with other methods and how to determine the accuracy of the results. The method survey part of the chapter also served as an orientation to the current state of the use of machine learning in strategy and management, complementing previous reviews of the rapidly evolving business intelligence domain. I acknowledge that several types of ML methods, including neural networks, graph-based techniques, decision maps, and Bayesian networks, among others, had to be omitted to attain a minimal depth in the discussion of the other methods, especially those based on NLP, classification, and clustering. Methods based on networks deserve an entirely separate treatment.

REFERENCES

Albright, R., Cox, J. & Daly, K. 2001. Skinning the cat: Comparing alternative text mining algorithms for categorization. In ***Proceedings of the 2nd Data Mining Conference of DiaMondSUG***. Chicago, IL. DM Paper (Vol. 113).

Allen, F., & Karjalainen, R. 1999. Using genetic algorithms to find technical trading rules. ***Journal of Financial Economics,*** 51(2): 245–271.

Alpaydin, E. 2014. ***Introduction to machine learning***. Cambridge, MA: MIT Press.

Arts, S., Cassiman, B., & Gomez, J.C. 2017. Text matching to measure patent similarity. ***Strategic Management Journal,*** 39(1): 62-84.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y.W. 2009. On smoothing and inference for topic models. In ***Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence***. Montreal, Canada: AUAI Press, 27-34.

Axtell, R.L. 2001. Zipf distribution of US firm sizes. ***Science,*** 293(5536): 1818–1820.

Baccianella, S., Esuli, A., & Sebastiani, F. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. ***Proceedings of LREC 2010***: 2200–2204.

Balazs, J.A., & Velásquez, J.D. 2016. Opinion mining and information fusion: A survey. ***Information Fusion,*** 27: 95–110.

Balsmeier, B., Li, G.C., Chesebro, T., Zang, G., Fierro, G., Johnson, K., Kaulagi, A., Lück, S., O'Reagan, D., Yeh, B., & Fleming, L. 2016. Machine learning and natural language processing on the patent corpus: data, tools, and new measures. ***UC Berkeley Working Paper***, UC Berkeley Fung Institute, Berkeley, CA.

Bandiera, O., Hansen, S., Prat, A., & Sadun, R. 2016. CEO behavior and firm performance. ***HBS Working Paper*** 17-083, Harvard Business School, Boston, MA. Available at: https://dash.harvard.edu/handle/1/30838134.

Bettis, R., Gambardella, A., Helfat, C., & Mitchell, W. 2014. Quantitative empirical analysis in strategic management. ***Strategic Management Journal,*** **35**(7): 949–953.

Bird, S., Klein, E., & Loper, E. 2009. Learning to classify text. In S. Bird, E. Klein, & E. Loper (Eds), ***Natural language processing with Python: Analyzing text with the natural language toolkit***, pp. 221–259. Sebastopol, CA: O'Reilly Media, Inc.

Bloomfield, R. 2008. Discussion of "annual report readability, current earnings, and earnings persistence." ***Journal of Accounting and Economics,*** 45(2): 248–252.

Blei, D.M., Ng, A.Y., & Jordan, M.I. 2003. Latent Dirichlet Allocation. ***Journal of Machine Learning Research,*** 3(Jan): 993-1022.

Bollen, J., Mao, H., & Zeng, X. 2011. Twitter mood predicts the stock market. ***Journal of Computational Science,*** 2(1): 1–8.

Brink, H., Richards, J., & Fetherolf, M. 2014. ***Real-world machine learning***. Shelter Island, NY: Manning.

Cavusgil, S.T., Kiyak, T., & Yeniyurt, S. 2004. Complementary approaches to preliminary foreign market opportunity assessment: Country clustering and country ranking. ***Industrial Marketing Management,*** 33(7): 607–617.

Chen, H., Chiang, R. H. L., & Storey, V.C. 2012. Business intelligence and analytics: From big data to big impact. ***MIS Quarterly***, 36(4): 1165-1188.

Chiu, C.Y., Chen, Y.F., Kuo, I.T., & Ku, H.C. 2009. An intelligent market segmentation system using k-means and particle swarm optimization. ***Expert Systems with Applications,*** 36(3): 4558–4565.

Choi, B., Poon, S.K., & Davis, J.G. 2008. Effects of knowledge management strategy on organizational performance: A complementarity theory-based approach. ***Omega,*** 36(2): 235–251.

Cubiles-De-La-Vega, M.-D., Blanco-Oliver, A., Pino-Mejías, R., & Lara-Rubio, J. 2013. Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques. ***Expert Systems with Applications***, 40(17): 6910–6917.

Debaere, P., Lee, H., & Lee, J. 2010. It matters where you go. Outward foreign direct investment and multinational employment growth at home. ***Journal of Development Economics***, 91(2): 301–309.

Devitt, A., & Ahmad, K. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In ***Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,*** Vol. 7. Prague, Czech Republic, 1–8.

Dinu, L., & Iuga, I. 2012. The Naïve-Bayes classifier in opinion mining: In search of the best feature set. In ***International Conference on Intelligent Text Processing and Computational Linguistics***. Springer, Berlin, 556–567.

Eggers, J.P., & Kaplan, S. 2009. Cognition and renewal: Comparing CEO and organizational effects on incumbent adaptation to technical change. ***Organization Science,*** 20(2): 461-477.

Elenkov, D.S., Judge, W., & Wright, P. 2005. Strategic leadership and executive innovation influence: An international multi-cluster comparative study. ***Strategic Management Journal,*** 26(7): 665–682.

Elfenbein, D.W., Knott, A.M., & Croson, R. 2017. Equity stakes and exit: An experimental approach to decomposing exit delay. ***Strategic Management Journal,*** 38(2): 278-299.

Elith, J., Leathwick, J.R., & Hastie, T. 2008. A working guide to boosted regression trees. ***Journal of Animal Ecology,*** 77(4): 1365–2656.

Ertek, G., Tapuku, D., & Arin, I. 2013. Text mining with RapidMiner. In M. Hofmann & R. Klinkenberg (Eds.), ***RapidMiner: Data mining use cases and business analytics applications***, 241–288. Boca Raton, FL: Chapman and Hall/CRC.

Friedman, J., Hastie, T., & Tibshirani, R. 2001. ***The elements of statistical learning***. Berlin: Springer.

Fujiwara, Y. 2004. Zipf law in firms bankruptcy. ***Physica A: Statistical Mechanics and its Applications,*** 337(1): 219–230.

Gabaix, X. 1999. Zipf's law for cities: An explanation. ***Quarterly Journal of Economics,*** 114(3): 739–767.

Gamache, D.L., McNamara, G., Mannor, M.J., & Johnson, R.E. 2015. Motivated to acquire? The impact of CEO regulatory focus on firm acquisitions. ***Academy of Management Journal,*** 58(4): 1261-1282.

Gans, J.S., Goldfarb, A., Lederman, M. 2017. Exit, tweets and loyalty. ***NBER Working Paper*** No. 23046. National Bureau of Economic Research, Cambridge, MA. Available at: http://www.nber.org/papers/w23046.

Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., & Heck, L. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. ***arXiv preprint***:1602.06291.

Glaeser, E.L., Kallal, H.D., Scheinkman, J.A., & Shleifer, A. 1992. Growth in cities. ***Journal of Political Economy,*** 100(6): 1126–1152.

Gow, I.D., Kaplan, S.N., Larcker, D.F., & Zakolyukina, A.A. 2016. CEO personality and firm policies. ***NBER Working Paper*** No. 22435. National Bureau of Economic Research, Cambridge, MA. Available at: http://www.nber.org/papers/w22435.

Hall, B.H., Jaffe, A.B., & Trajtenberg, M. 2001. The NBER patent citation data file: Lessons, insights and methodological tools. ***NBER Working Paper*** No. 8498. National Bureau of Economic Research, Cambridge, MA. Available at: http://www.nber.org/papers/w8498.

Hall, P., Dean, J., Kabul, I.K., & Silva, J. 2014. An overview of machine learning with SAS Enterprise Miner. In ***Proceedings of the SAS Global Forum 2014 Conference***. Available at: https://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf.

Hoberg, G., & Phillips, G. 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. ***Review of Financial Studies,*** 23(10): 3773–3811.

Hoffman, M., Bach, F.R., & Blei, D.M. 2010. Online learning for Latent Dirichlet Allocation. In ***Advances in Neural Information Processing Systems,*** 23: 856-864.

Hu, D., Zhao, J.L., Hua, Z., & Wong, M.C.S. 2012. Network-based modeling and analysis of systemic risk in banking systems. ***MIS Quarterly,*** 36(4): 1269-1291.

Hu, M., & Liu, B. 2004. Mining and summarizing customer reviews. In ***Proceedings of the 10thACM SIGKDD International Conference on Knowledge Discovery and Data Mining***, Seattle, WA, 168–177. Available at: http://dl.acm.org/citation.cfm?id=1014052&picked=prox.

Huang, A. 2008. Similarity measures for text document clustering. In ***Proceedings of the 6thNew Zealand Computer Science Research Student Conference*** (NZCSRSC2008), Christchurch, New Zealand, 49–56.

Jordan, M.I., & Mitchell, T.M. 2015. Machine learning: Trends, perspectives, and prospects. ***Science,*** 349(6245): 255–260.

Kanze, D., Huang, L., Conley, M.A., & Higgins, E.T. 2017. We ask men to win & women not to lose: Closing the gender gap in startup funding. ***Academy of Management Journal***, preprint amj-2016.

Kaplan, S. 2012. Identifying breakthroughs: Using topic modeling to distinguish the cognitive from the economic. ***Academy of Management Proceedings,*** 2012(1).

Kotu, V., & Deshpande, B. 2014. ***Predictive analytics and data mining: Concepts and practice with Rapidminer***. Waltham, MA: Morgan Kaufmann.

Kučera, H., & Francis, W.N. 1967. ***Computational analysis of present-day American English***. Providence, RI: Brown University Press.

Lamberg, J.A., Tikkanen, H., Nokelainen, T., & Suur-Inkeroinen , H. 2009. Competitive dynamics, strategic consistency, and organizational survival. ***Strategic Management Journal,*** 30(1): 45–60.

Lange, D., Boivie, S., & Westphal, J.D. 2015. Predicting organizational identification at the CEO level. ***Strategic Management Journal,*** 36(8): 1224-1244.

Lau, R.Y.K., Liao, S.S.Y., Wong, K.F., & Chiu, D.K.W. 2012. Web 2.0 environmental scanning and adaptive decision support. ***MIS Quarterly,*** 36(4): 1239-1268.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. 2014. The parable of Google Flu: Traps in big data analysis. ***Science,*** 343(6176): 1203–1205.

Lee, C., Shleifer, A., & Thaler, R. 1991. Investor sentiment and the closed-end fund puzzle. ***The Journal of Finance,*** 46(1): 75–109.

Lee, P.M., & James, E.H. 2007. She'-e- os: Gender effects and investor reactions to the announcements of top executive appointments. ***Strategic Management Journal,*** 28(3): 227-241.

Leech, G. 1991. The state of the art in corpus linguistics. In J. Svartvik, K. Aijmer, & B. Altenberg (Eds.), ***English corpus linguistics: Studies in honour of Jan Svartvik,*** pp. 8–29. London: Longman.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. ***Journal of Accounting and Economics,*** 45(2): 221–247.

Li, F. 2010a. Textual analysis of corporate disclosures: A survey of the literature. ***Journal of Accounting Literature,*** 29: 143–165.

Li, F. 2010b. The information content of forward-looking statements in corporate filings —a naïve Bayesian machine learning approach. ***Journal of Accounting Research,*** 48(5): 1049–1102.

Li, G.C., Lai, R., D'Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Amy, Z.Y., & Fleming, L. 2014. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). ***Research Policy,*** 43(6): 941–955.

Liu, L.Y., Jiang, T.J., & Zhang, L. Hashtag recommendation with topical attention-based LSTM. 2016. In ***Proceedings of the 26th International Conference on Computational Linguistics***, Osaka, Japan, 943-952.

Liab, H., & Sun, J. 2011. Principal component case-based reasoning ensemble for business failure prediction. ***Information & Management,*** 48(6): 220-227.

Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science: Irvine, CA.

Loughran, T., & McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. ***The Journal of Finance,*** 66(1): 35–65.

Loughran, T., & McDonald, B. 2014. Measuring readability in financial disclosures. ***The Journal of Finance,*** 69(4): 1643–1671.

Loughran, T, & McDonald, B. 2016. Textual analysis in accounting and finance: A survey. ***Journal of Accounting Research,*** 54(4): 1187–1230.

Louridas, P., & Ebert, C. 2017. Machine learning. ***Computing Edge***, April 2017: 8–13.

Lugmayr, A. 2013. Predicting the future of investor sentiment with social media in stock exchange investments: A basic framework for the DAX Performance Index. In M. Friedrichsen & W. Mühl-Benninghaus (Eds.), ***Handbook of social media management***, 565–589. Berlin, Heidelberg: Springer Media Business and Innovation.

Magnussen, S., Tomppo, E., & McRoberts, R.E. 2010. A model-assisted k-nearest neighbour approach to remove extrapolation bias. ***Scandinavian Journal of Forest Research,*** 25(2): 74 – 184.

Manning, C.D., & Schütze, H. 1999. Topics in information retrieval. In ***Foundations of statistical natural language processing,*** pp. 539–554. Cambridge, MA: MIT Press.

Manning, C.D., Raghavan, P., & Schütze, H. 2008. ***Introduction to information retrieval***. Cambridge, UK: Cambridge University Press.

Marco, A.C., Tesfayesus, A., & Toole, A.A. 2017. Patent Litigation Data from US District Court Electronic Records (1963–2015). ***USPTO Economic Working Paper*** No. 2017-06. Available at SSRN: https://ssrn.com/abstract=2942295.

Marcus, M.P., Marcinkiewicz, M.A., & Santorini, B. 1993. Building a large annotated corpus of English: The Penn Treebank. ***Computational Linguistics,*** 19(2): 313–330.

Marsland, S. 2015. Learning with trees. In ***Machine learning: An algorithmic perspective***, pp. 249–266. Boca Raton, FL: CRC Press .

Matej, M., Miroslav, P. 2013. Medical data mining. In M. Hofmann & R. Klinkenberg (Eds.), ***RapidMiner: Data mining use cases and business analytics applications***, pp. 241–288. Boca Raton, FL: Chapman and Hall/CRC.

Melville, P., Gryc, W., & Lawrence, RD. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In ***Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***, Paris, France, 1275–1284.

Metzler, D., Dumais, S., & Meek, C. 2007. Similarity measures for short segments of text. In G. Amati, C. Carpineto, & G. Romano (Eds), ***Advances in information retrieval***. ECIR 2007. Lecture Notes in Computer Science, Vol 4425, pp. 16-27. Berlin, Heidelberg: Springer.

Mostafa, M. M. 2013. More than words: Social networks' text mining for consumer brand sentiments. ***Expert Systems with Applications,*** 40(10): 4241–4251.

Nadkarni, S., & Barr, P. S. 2008. Environmental context, managerial cognition, and strategic action: An integrated view. ***Strategic Management Journal,*** 29(13): 1395-1427.

Ngai, E. W., Xiu, L., & Chau, D.C. 2009. Application of data mining techniques in customer relationship management: A literature review and classification. ***Expert Systems with Applications,*** 36(2): 2592–2602.

Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D., & Joksimovic, I. 2013. The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. ***Expert Systems with Applications,*** 40(15): 5932–5944.

Pang, B., & Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In ***Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics***, Barcelona, Spain, 271.

Pang, B., & Lee, L. 2008. Opinion mining and sentiment analysis. ***Foundations and Trends in Information Retrieval*** 2(1–2): 1–135.

Piryani, R., Madhavi, D., & Singh, V.K. 2017. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. ***Information Processing & Management,*** 53(1): 122–150.

Porter, M.F. 1980. An algorithm for suffix stripping. ***Program,*** 14(3): 130–137.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. 2008. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In ***Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data ining***, Las Vegas, Nevada, 569-577.

Punj, G., & Stewart, D.W. 1983. Cluster analysis in marketing research: Review and suggestions for application. ***Journal of Marketing Research,*** 20: 134–148.

Pustejovsky, J., & Stubbs, A. 2012. Corpus analytics. In J. Pustejovsky & A. Stubbs (Eds.), ***Natural language annotation for machine learning***, pp. 53–65. Sebastopol, CA: O'Reilly Media, Inc.

Quinlan, J.R. 1979. Discovering rules by induction from large collections of examples. In D. Michie (Ed.), ***Expert Systems in the Microelectronic Age***, 168–201. Edinburgh, UK: Edinburgh University Press.

Ravi, K., & Ravi, V. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. ***Knowledge-Based Systems,*** 89: 14-46.

Roth, K. 1992. International configuration and coordination archetypes for medium-sized firms in global industries. ***Journal of International Business Studies,*** 23(3): 533–549.

Salton, G., & Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. ***Information Processing & Management,*** 24(5): 513–523.

Salton, G. 1991. Developments in automatic text retrieval. ***Science*** 253(5023): 974–980.

Savage, N. 2012. Gaining wisdom from crowds. ***Communications of the ACM,*** 55(3): 13–15.

Schaffer, C.M., & Green, P.E. 1998. Cluster-based market segmentation: Some further comparisons of alternative approaches. ***Journal of the Market Research Society,*** 40(2): 155-164.

Sebastiani, F. 2002. Machine learning in automated text categorization. ***ACM Computing Surveys,*** (CSUR) 34(1): 1–47.

Shannon, C.E. 1948. A mathematical theory of communication. ***Bell System Technical Journal,*** 27: 379–423.

Shannon, C.E. 1951. Prediction and entropy of printed English. ***Bell System Technical Journal,*** 30: 50–64.

Siegel, E. 2016. ***Predictive analytics: The power to predict who will click, buy, lie, or die,*** pp. 103-110. Hoboken, NJ: Wiley.

Singh, J., Verbeke, W., & Rhoads, G.K. 1996. Do organizational practices matter in role stress processes? A study of direct and moderating effects for marketing-oriented boundary spanners. ***Journal of Marketing***, 60(3): 69-86.

Slack, F.J., Orife, J.N., & Anderson, F.P. 2010. Effects of commitment to corporate vision on employee satisfaction with their organization: An empirical study in the United States. ***International Journal of Management,*** 27(3): 421-436.

Sohn, M.H., You, T., Lee, S.L., & Lee, H. 2003. Corporate strategies, environmental forces, and performance measures: A weighting decision support system using the k-nearest neighbor technique. ***Expert Systems with Applications,*** 25(3): 279-292.

Sohn, S.Y., & Kim, J.W. 2012. Decision tree-based technology credit scoring for start-up firms: Korean case. ***Expert Systems with Applications***, 39(4): 4007-4012.

Srivastava, S.B., Goldberg, A., Manian, V.G., & Potts, C. 2017. Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. ***Management Science***. (Articles in Advance, doi:10.1287/mnsc.2016.2671).

Stock, G.N., Greis, N.P., & Kasarda, J.D. 2000. Enterprise logistics and supply chain structure: The role of fit. ***Journal of Operations Management***, 18(5): 531–547.

Struhl, S. 2013. ***Market segmentation***. Chicago, IL: American Marketing Association Press.

Struhl, S. 2015. In the mood for sentiment. In ***Practical text analytics: Interpreting text and unstructured data for business intelligence***. London, UK: Kogan Page Publishers: 120–143.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon-based methods for sentiment analysis. ***Computational Linguistics,*** 37(2): 267–307.

Tam, K.Y., & Kiang, M.Y. 1992. Managerial applications of neural networks: The case of bank failure predictions. ***Management Science,*** 38(7): 926 – 947.

Tan, T. Z., Quek, C., & Ng, G.S. 2007. Biological brain-inspired genetic complementary learning for stock market and bank failure prediction. ***Computational Intelligence,*** 23(2): 236-261.

Teh, Y., Jordan, M., Beal, M., & Blei, D. 2007. Hierarchical Dirichlet processes. ***Journal of the American Statistical Association,*** 101(476): 1566–1581.

Tetlock, P.C. 2007. Giving content to investor sentiment: The role of media in the stock market. ***The Journal of Finance,*** 62(3): 1139–1168.

Trajtenberg, M., Shiff, G., & Melamed, R. 2006. The "names game": Harnessing inventors' patent data for economic research. ***NBER Working Paper*** No. w12479. National Bureau of Economic Research, Cambridge, MA. Available at: http://www.nber.org/papers/w12479.

Tsai, C.F., & Chen, M.L. 2010. Credit rating by hybrid machine learning techniques. ***Applied Soft Computing,*** 10(2): 374–380.

Tsytsarau, M., Palpanas, T. 2012. Survey on mining subjective data on the web. ***Data Mining and Knowledge Discovery,*** 24(3): 478–514.

Turing, A.M. 1950. Computing machinery and intelligence. ***Mind,*** 59: 433–460.

Van de Vrande, V., De Jong, J.P., Vanhaverbeke, W., & De Rochemont, M. 2009. Open innovation in SMEs: Trends, motives and management challenges. ***Technovation,*** 29(6): 423–437.

Varian, H.R. 2014. Big data: New tricks for econometrics. ***Journal of Economic Perspectives,*** 28(2): 3–28.

Wang, C.H. 2009. Outlier identification and market segmentation using kernel-based clustering techniques. ***Expert Systems with Applications,*** 36(2): 3744–3750.

Wang, C., Paisley, J., & Blei, D.M. 2011. Online variational inference for the Hierarchical Dirichlet Process. In ***Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)***, Fort Lauderdale, Florida, 752-760.

Wang, G., Ma, J., Huang, L., & Xu, K. 2012. Two credit scoring models based on dual strategy ensemble trees. ***Knowledge-Based Systems,*** 26(2): 61-68.

Wedel, M., & Kamakura, WA. 2012. ***Market segmentation: Conceptual and methodological foundations*** (Vol. 8). New York: Springer Science & Business Media.

Westreich, D., Lessler, J., & Funk, M.J. 2010. Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. ***Journal of Clinical Epidemiology,*** 63(8): 826–833.

Willett, P. 1988. Recent trends in hierarchic document clustering: A critical review. ***Information Processing & Management,*** 24(5): 577–597.

Williams, C., & Lee, S.H. 2009. Resource allocations, knowledge network characteristics and entrepreneurial orientation of multinational corporations. ***Research Policy,*** 38(8): 1376–1387.

Yeh, I.C., Lien, C.H. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. ***Expert Systems with Applications,*** 36(2): 2473-2480.

Younge, K.A., & Kuhn, J.M. 2015. Patent-to-patent similarity: A vector space model. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2709238.

Yu, Y., Duan, W., & Cao, Q. 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. ***Decision Support Systems,*** 55(4): 919–926.

Xiao, H., & Stibor, T. 2010, October. Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In ***Proceedings of 2nd Asian Conference on Machine Learning***, Tokyo, Japan, 63-78.

APPENDIX – CHAPTER 2

Sentiment Analysis Applied to Competitors

The corporate motto fulfills essential roles for the firm: attracting customers, distinguishing the firm from competitors, signaling the core of the firm's culture, and motivating employees. Mottos are essential in commerce, have been in use for hundreds of years, and can be trademarked as part of the firm's brand. Firm mottos are designed to evoke emotion – in the customer, the employee, and the competitor – and as a result, linguistic properties related to sentiment and degree of subjectivity are useful in classifying them. Firms that evoke similar sentiments as found in the text analysis may be competing for a particular type of customer and may be closer in competition than others with very different corporate mottos; this may aid the researcher with a layer of classification beyond that of industry and location. For this example, I chose again a freely available dataset and a set of intuitive tools. The dataset is a listing of corporate mottos used by banks and is collected by The Financial Brand and available online.[21] The version of the data current as of the writing of this appendix and used for this analysis contains 888 financial firms. The analysis was done in RapidMiner with two different NLP sentiment analysis packages, AYLIEN and the Meaning Cloud Sentiment Analyzer. The misclassification error was lower for this data using the latter package; I present the steps and output from the Meaning Cloud package. Both packages run on the cloud and so require registration for a free account. Implementing this in Python NLTK is of course possible, though doing so will require a significant implementation effort. The schematic for the process is in Figure A.1:

[21] "The Biggest List of Financial Slogans Ever," The Financial Brand, https://thefinancialbrand.com/1779/financial-slogans/. Accessed September 30, 2017.

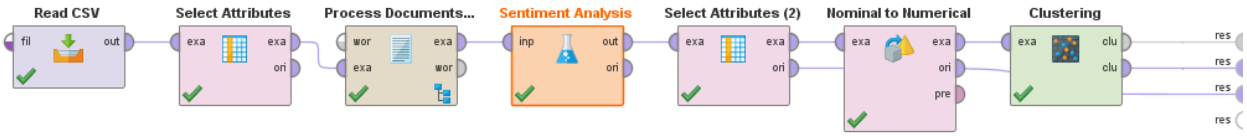

Figure A.1. Implementation of a sentiment analyzer and clustering based on sentiment characteristics of text; the clustering step is optional – the output of the Meaning Cloud sentiment analysis operator is a vector of characteristics (polarity, on a five-point scale ranging from very negative to very positive; objectivity taking values of objective or subjective; irony taking values of ironic or not ironic; agreement taking a binary value as well; and finally the confidence of the classification on a scale of 0 to 100).

Row No.	FirmName	cluster	text	polarity(text) = NONE	polarity(text) = P+	polarity(text) = P	polarity(text) = N+
1	121 Financial Credit Union	cluster_9	banking focused on you	1	0	0	0
2	1st Advantage Bank	cluster_9	expertise you need service you deserve	0	1	0	0
3	1st Financial Federal Credit Union	cluster_0	the better way to bank	0	0	1	0
4	1st Mariner Bank	cluster_0	we built this bank for you	0	0	1	0
5	1st National Bank	cluster_9	grow with us at home with first	1	0	0	0
6	1st National Bank of So. Florida	cluster_9	your first choice	1	0	0	0
7	1st Source Bank	cluster_9	your partners from the first	1	0	0	0
8	A+ Federal Credit Union	cluster_9	with you every step	1	0	0	0
9	ABN AMRO Bank	cluster_9	making more possible	1	0	0	0
10	ABNB Federal Credit Union	cluster_0	open honest hardworking	0	0	1	0
11	AIG	cluster_0	the strength to be there we know money	0	0	1	0
12	ANZ	cluster_5	the better we know you the more we can do	0	1	0	0
13	ASB Bank	cluster_0	creating futures	0	0	1	0
14	Abbey National Bank	cluster_6	get the abbey habit turning banking on its head...	0	0	0	1
15	Abington Bank	cluster_0	banking for people with better things to do	0	0	1	0
16	Absa Bank	cluster_9	today tomorrow together	1	0	0	0
17	Access National Bank	cluster_9	the difference is access	1	0	0	0
18	Achieva Credit Union	cluster_0	dream it achieve it	0	0	1	0
19	Acru	cluster_9	money life	1	0	0	0
20	Addison Avenue FCU	cluster_0	we listen you prosper	0	0	1	0
21	Advantage Plus FCU	cluster_4	not for profit for people	0	0	0	1
22	Afena Credit Union	cluster_9	we are already there	1	0	0	0
23	Affinity Group Credit Union	cluster_9	changing lives one member at a time	1	0	0	0
24	Affinity Plus FCU	cluster_4	not for profit for people	0	0	0	1

Figure A.2. RapidMiner output (partial) showing some of the text sentiment categories and the cluster assignments for a sample of the firms.

Cluster Model

```
Cluster 0: 318 items
Cluster 1: 1 items
Cluster 2: 3 items
Cluster 3: 4 items
Cluster 4: 4 items
Cluster 5: 14 items
Cluster 6: 1 items
Cluster 7: 5 items
Cluster 8: 1 items
Cluster 9: 537 items
Total number of items: 888
```

Figure A.3. Output of ten clusters based on the sentiment analysis vectors. Cluster 9 is dominated by firms with no sentiment polarity; cluster 0 by firms with positive sentiment found in the mottos.

A sample of the output is in Figure A.2, and the summary view of the clusters in Figure A.3. The confidence in the classification is above 99%. Only 15 firms had a negative message, whereas 376 firms had positive messages. Of the 888 texts, 177 were classified as subjective; no firms had irony in their mottos.

This is a very simple example of how one may use sentiment analysis to determine firms with similar strategies; of course, the text analysis portion would need to be supplemented by other data and other methods. However, the steps outlined here can easily be extended to other data and other questions. This appendix serves as a simple example available with free tools and datasets. The UCI ML repository is an excellent source of additional examples, papers, code and data, and several of the machine learning books outlined in the references point to additional examples and online sources for the interested reader.

Chapter 3

The Need for Speed: Effects of Uncertainty Reduction in Patenting

Mike H. M. Teodorescu

ABSTRACT

Patents are essential in commerce to establish property rights for ideas and to give equal protection to firms that develop new technologies. Startups depend on the protection of intellectual property to bring a product from concept to market. However, the market for technology ideas has been recognized as an inefficient market in the management and economics literatures. While information asymmetry and expropriation risks have been studied extensively, the question of the effects of prepatent grant uncertainty on firm outcomes remains open. This paper introduces a novel analysis based on internal US Patent and Trademark Office databases, exploiting an exogenous shock to startup firms from a previously unstudied executive action involving reduction of patent pendency (time from application to patent decision) for green technology patents. The findings are that treated startups have greatly increased sales, employment, and venture funding. The paper also introduces a novel method for constructing a control group using a classification algorithm rooted in natural language processing, which can be used in conjunction with traditional econometric approaches such as difference-in differences analysis beyond the topic of this paper.

KEYWORDS: Innovation; Accelerated Patenting; Green Technology; Textual Similarity; Classifier.

INTRODUCTION

The need to provide a legal instrument to protect technologies was recognized as vital by the Founding Fathers of the United States. In fact, the concept of a patent is enshrined in the US Constitution. A patent is a property title for a technologically useful and novel idea and constitutes a right to manufacture a product with the idea, license it, and sell it. It is due to this legal protection that startups with no other assets can break into a market or generate their own market. In the absence of strong property rights protections, each interaction in which the invention is shared carries a significant risk of expropriation. However, given the current intellectual property (IP) legal framework in the US, once a firm is given a title to its technology idea through the legal instrument of the patent, it may use that title to generate revenue or obtain funding.

However, asymmetry of information (Arrow, 1962), risk of expropriation (Anton & Yao, 1994 and 2004), and uncertainty (Arora & Gambardella, 2010) are all barriers to an efficient market for technology ideas. Simply put, it is difficult to evaluate whether a technology is likely to be valuable or to obtain patent protection. Arora and Gambardella (2010) point out that evaluating a technology's potential and technical viability is inherently hard and is a source of uncertainty for firms. Uncertainty in a startup's technology can be a particular deterrent to investors and can ultimately harm the firm's growth and survival prospects. This paper focuses on a little-studied source of uncertainty: the amount of time necessary to obtain a decision on the patent application, also known as the *patent pendency term*.

Patent pendency is largely a function of complexities in the patent system. It is outside the control of the inventors, attorneys, and other stakeholders, such as investors. The average patent application pendency (time from application to grant) in the decade prior to 2008 was close to three years on average, with a standard deviation of over a year and a half. Startups in particular

are vulnerable to such high variations in timing, as often the patent application is the only asset they can use to raise funding or generate revenue. The complexities of the patenting process are explained in the institutional context and literature review sections that follow.

To analyze the effects on startups of a reduction in pendency term uncertainty, I look at an exogenous shock to firms generated by a program the US Patent and Trademark Office (USPTO) ran from 2009-2012 that prioritized patents in categories of inventions deemed of national interest: the Green Technology Pilot Program, which accelerated patent applications pertaining to environmental protection, reduction of energy consumption, and clean energy generation.

I construct a novel dataset using internal USPTO databases through a yearlong data collection process conducted in collaboration with the USPTO. I also interviewed officials from across the USPTO to obtain institutional details on matters such as the examination processes and assignment to the Green Technology Program. With access to the internal databases of the patent office, I analyze all of the other programs that enable inventors to be examined sooner, the "accelerated examination programs," and determine that the Green Technology Pilot Program is the most appropriate for studying the effects of patent pendency uncertainty on startup outcomes, as it has more data than the other accelerated examination avenues and does not suffer from selection biases based on patent quality or high financial entry costs. The novelty of the setup makes this the first study of the effects of patent term reduction on firm outcomes.

The paper follows a difference-in-differences econometric approach using the treatment and control groups from the USPTO program, where patents accepted into the program are matched to startups. A method for constructing a control group is proposed here as well, based on constructs from the field of natural language processing, and tailored to the nature of patent texts. A second analysis is performed using this text-matched group, and a third analysis is performed using

coarsened exact matching. Extensions of the text-matching method are proposed for fields beyond patents.

I find that the treated firms fare significantly better than firms that do not benefit from an accelerated patent examination process. Firms sell over 30% more, hire more employees (again double digit percentage increases), and raise more venture capital (double digit percentage increases compared to firms without patent acceleration). The significance of the findings is twofold: startups benefit significantly from reducing the uncertainty in the time necessary to evaluate their patent applications, and governments can encourage growth in certain sectors by using the levers of IP policy.

The following section provides a review of the literature and an in-depth view of the institutional context. The subsequent sections give the data description, methodology, results, robustness checks, and conclusions.

LITERATURE REVIEW

The patent examination process (detailed in the institutional context section of this paper) involves multiple steps and takes a variable amount of time from application to application, in some cases extending over three or more years. This paper focuses on an under-studied source of uncertainty, namely patent grant delay. The relevant measure here is patent pendency – the amount of time a patent application is under examination in the patent office, specifically the time between the filing of a patent application and its decision date.

By clarifying the ownership and merits of an idea, patents offer inventors an opportunity to reduce uncertainty and friction in the market for technology ideas. Uncertainty in the market for technology is a topic that requires more research, as the majority of research into this market's

inefficiencies focuses on the impacts of information asymmetry (Arora & Gambardella, 2010: 657). Studies by Gans *et al.* (2008) and Hegde and Luo (2018) show that the uncertainty generated by patent grant delays affects the technology licensing market. For instance, Gans *et al.* (2008) find that a patent grant substantially increases the likelihood of selling a license to the technology.

A patent application is the upper bound of what may ultimately become a patent, and thus "significant uncertainty exists over the scope of the patent rights ultimately allowed and the enforceability of allowed claims through litigation" (Gans *et al.,* 2008, p. 984). Gans *et al.* (2008) distinguish between different types of uncertainty while focusing on its effects on licensing: uncertainty in patent applications; uncertainty in patent pendency (the amount of time necessary to obtain a decision on the patent); uncertainty in the patent grant (whether the inventor will prevail in demonstrating the merits of the invention); and post-patent enforcement uncertainty (whether the patent will be enforceable once granted). The latter is a fruitful area of research and has generated the patent litigation literature, which shows that there is significant heterogeneity in ability to enforce patents across different technology classes (Lerner, 1995), that failure to defend a patent may affect subsequent firm innovation and exit decisions (Galasso & Shankerman, 2015), and that enforcement uncertainty may even lead to abnormal market returns (Sidak & Skog, 2015).

Hegde and Luo (2018) focus on the reduction of pre-patent grant uncertainty following the passing of the American Inventors Protection Act (AIPA) in 1999, which mandated the publication of patent applications at 18 months; this reduced uncertainty for idea buyers, resulting in an increase in pre-grant licensing deals. Gans *et al.* (2008) also focus on licensing, finding that a reduction in the uncertainty about the granting of a patent itself increases licensing deals. The various sources of uncertainty can be seen in Figure 1. This paper focuses on the post-publication decision time T_2, while the bulk of the literature focuses on the Publication and Litigation timing.

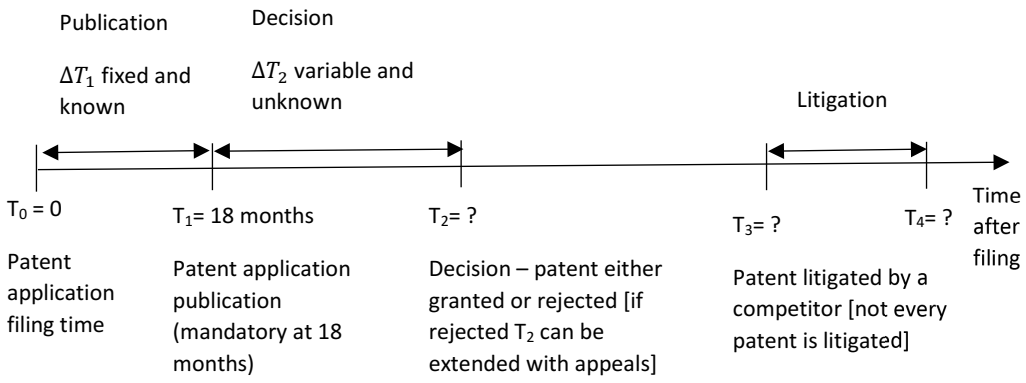

Figure 1. Sources of uncertainty in the patenting process for firms that patent. The time from patent application filing to a decision by the patent examiner is unknown and is the subject of this paper; the outcome of the patent application is also a source of uncertainty for the applicant and potential investors; and finally, should the patent be litigated after its issuance, the timing of a final court disposition is also a source of uncertainty for the firm that applied for the patent and for its investors and customers.

In related work, Farre-Mensa *et al.* (2016) find that reducing uncertainty about the validity of a patent application and its scope helps small inventors' IPOs and facilitates their access to capital; their construction is based on an IV using quasi-random assignment of applications to patent examiners. Hsu and Ziedonis (2011) find that patent grants are more important for inventors with little initial reputation (thus also focusing on the reduction of uncertainty about the patent application's scope and viability). Sukhatme and Cramer (2014) find that extensions in patent term post-TRIPS (1994) are valuable for inventors, resulting in higher market returns. This finding reinforces the consensus that *post patent grant,* the frictions in the market for ideas are reduced, as much of the uncertainty regarding the scope of the invention is resolved. The following section describes in depth the institution of the USPTO and the patent examination process, and gives details on the Green Technology Program used as a shock in this paper.

INSTITUTIONAL CONTEXT

The US Patent and Trademark Office has a consecrated role in the US Constitution: "To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries" (Article 1, Section 8). Since the writing of the Constitution, the role of the Patent Office as a grantor of time-limited monopolies[22] for inventions has been regulated by acts of Congress. The Patent Act of 1952, codified in 35 U.S.C., established the modern form of the USPTO as a division of the Department of Commerce of the executive branch. Two other major changes to the legal framework of the USPTO were made through the American Inventors Protection Act (AIPA, passed 1999, updated 2002), which required all patent applications to be published 18 months after filing, allowed for patent term adjustments in cases of delay caused by patent office backlog, and established the Request for Continued Examination procedure to permit an application to be revived after final rejection by an examiner; and the America Invents Act (2012), which aligned the US with other patent systems by creating a "first to file" system (out of two identical inventions, the first inventor to file is given the patent) and established an expedient administrative court to judge post-grant matters called the Patent Trial and Appeals Court (PTAB).

Apart from these major legal framework changes, the USPTO governs itself through executive actions that can be issued by the Director of the USPTO as the Under Secretary of Commerce. A fully fee-funded agency (with fees collected from patent and trademark applicants and owners), the USPTO has significant latitude in terms of the executive actions that affect innovation in the US. The USPTO can, for example, institute pilot programs that streamline application processes

[22] As regulated by Congress for the patent term, the amount of time a patent is valid as long as maintenance fees are paid, currently a term of 20 years.

for certain categories of inventors or for certain topics through executive actions, or change internal application review processes or employee incentives, all of which can have an impact on inventors and the IP environment in general. This paper uses the executive action passed in 2009, which amounted to an exogenous shock to the green technology startup firm community. Broadly, the authority for the Green Technology Program lies in the USPTO's legal right to decide that categories of inventions are of national interest.

Legal Framework

Under US Code 37 CFR 1.102, the USPTO may award *prioritized examination* to certain categories of applicants the Office or the executive branch deems of "particular importance to some branch of the public service" when "the head of some department of the Government requests immediate action for that reason" (37 CFR 1.102(b)). Prioritized examination (also known as "accelerated exam") allows certain categories of inventors or inventions to be examined out of turn (i.e., not in the order received). In practice, this involves an examination decision within 12 months of the granting of a petition for prioritized exam, reducing the uncertainty of the patent examination and making pendency predictable for those accepted for accelerated examination.

The processing time for a petition to accelerate an application is very short compared to the actual patent examination time. One interviewee commented, "we tried to process all of [the accelerated exam petitions] within 30 days" (interview with a petitions adjudicator in February 2017), a claim that is supported by my data. The USPTO runs several priority examination programs based on categories deemed by the executive to be of national interest, which have included environmental conservation, counterterrorism, and AIDS and cancer therapies. These programs require only a petition from the inventor requesting priority; *no fees or other application materials are required.* Unlike programs such as Track One and the regular Accelerated Exam, which require firms to pay

significant sums to be considered out of turn (meaning that those with limited resources, such as startups, may select only their best inventions for treatment), in the case of the Green Technology Program, the treatment is costless.

The Accelerated Exam, which is independent of the national interest tracks, exemplifies the selection issues with programs other than the Green Technology Program. The Accelerated Exam is a fee-based program that requires an extensive search report performed by the applicant, involving considerable legal costs. Patents in this program are thus of higher quality than the average patent, as significantly higher legal costs are paid upfront by the firm *before the examination* at the USPTO even begins.

The Green Technology Program was established in 2009 (see Federal Register dated December 8 2009: 64666) and ended in 2012. It is the largest program based on topics deemed high-priority (3,500 patents qualified). Based on interviews (in February and May of 2017) with patent petitions officers within the USPTO who administered the program and made decisions about priority, I learned that the only requirement for an application to be granted a priority exam through this program was that it be a form of green technology. To avoid potential incentive conflicts, the examiner assigned to the patent application was not the one to determine whether the patent application should be evaluated out of turn. This determination was made by a separate officer from a different section of the USPTO (usually the Office of Petitions). This officer evaluated whether the *claims* of the patent application were in the green technology field. The claims of the patent application are the legal instrument that defines the boundaries of the idea covered in the patent description and are the essence of the patent as seen in the court system.

Patent Examination Process

To illustrate the uncertainty in the regular patent examination process, let us review the steps in the examination. The patent application process includes a number of milestones: docketing and classification (involving the acceptance of a patent application, the assigning of an application number and a priority date – an important legal step that prevents another inventor from claiming the same subject matter in the future – and the assigning of one or several patent classes); the first office action (in which an examiner rules on the merits of the application and the patent claims, resulting in either a *non-final rejection* or, rarely, a *first action allowance* that grants the patent pending payment of fees); a second office action (either *allowance,* granting the patent pending payment of fees; or, most often, a *final rejection*, which can be appealed); and an appeal of the final rejection (resulting in another rejection or an allowance), which is a court process at the Patent Trial and Appeals Board.

A common alternative to the appeal of the final rejection is the Request for Continued Examination (RCE), a purely administrative process involving the same examiner that extends the review time and is renewable for several cycles if additional review results in another rejection. RCEs and appeals greatly extend the patent decision time. While they are mentioned here for sake of completeness, they are outside the purview of this study.

A separate outcome is abandonment, which can occur at any time prior to patent issuance and is the result of either non-payment of fees or failure to respond to a patent office communication within 6 months. *Patent pendency* is a metric that measures the time elapsed between the *patent filing date* and the *final decision date* (the final decision is either an allowance or a final rejection). Patent pendency is an important metric because longer patent pendency injects uncertainty into the innovation system as a whole (Gans *et al.*, 2008). Delays in patent grant decisions due to backlogs

in the patent examination system impede an efficient market for ideas (Arora & Gambardella, 2010; Hegde & Luo, 2018; Farre-Mensa *et al.*, 2016).

Pendency used to be three or more years, and is now slightly over two years,[23] still a substantial impediment to the sale or licensing of intellectual property and a source of uncertainty for investors evaluating startups. Table 1 shows the summary statistics of the patent pendency of all patents 1998-2016 and of all patents in 2008-2012 (the year before treatment and the period contemporary with the treatment); while the pendency itself is multiple years on average, notice the high standard deviation of over a year in both:

Table 1: Summary Statistics for Patent Pendency

	(mean)	(sd)	(min)	(max)
Pendency All Patents 1998-2016 (days)	1011.02	561.34	21	6270
Pendency All Patents 2008-2012 (days)	988.72	448.70	34	2995

While the existence of a "patent pending" designation on a product is an assertion of potential future litigation should an infringer copy the invention, the patent applicant does not have the right to litigate prior to a decision on the patent, though the existence of a published patent application may facilitate licensing agreements (Hegde & Luo, 2018). A patent application is, simply put, an indication of *the maximum protection the inventor may obtain*. Examiners may reject the patent's claims in their entirety or reduce their scope, information that will be available to the public only after a second office action. Thus, the outcome of a patent application and the merits of its claims are uncertain until after the second office action. This source of uncertainty has an impact on the investment perspectives of a startup.

[23] The USPTO considers patent pendency, both for the time of the first action and for the time of the final decision or second office action, to be an important metric of performance. The USPTO has been working on reducing backlog as a strategic goal (USPTO 2014-2018 Strategic Plan, www.uspto.gov/strategicptlan, reviewed in print form at the USPTO library in August 2016) and reports pendency as a monthly metric at the USPTO Data Visualization Center – Patents Dashboard https://www.uspto.gov/corda/dashboards/patents/kpis/kpiOverallPendency.kpixml, Accessed 2/21/2017.

DATA

The unique and new dataset constructed for this paper is the result of a yearlong research project within the USPTO, involving access to internal USPTO databases and research meetings and interviews with leaders in the USPTO, as well as interviews and conversations with patent examiners, classifiers, and petitions officers. A substantial effort was made to ensure the highest levels of data quality through the use of independent data paths to cross-check the validity of the sample, programmatic disambiguation of firm names, and the use of probabilistic matching techniques. In this section, I detail the data sources and some of the steps taken to construct the dataset.

Patent Data

This is a novel dataset in the management literature. It is derived from multiple sources, including internal data from the USPTO's databases, specifically the Image File Wrapper (IFW) and the Patent Application Location and Monitoring (PALM) system, and public sources such as USPTO's PatentsView, which provides disambiguated assignee and inventor names.[24] The IFW is the largest internal database of the USPTO. It contains scanned images of every page in every patent application file, including decisions, and assigns codes based on document type. While it is the most comprehensive database, the IFW is for the most part not machine readable, as it contains non-OCR-ed scanned images that go back to the beginning of the patent office (covering every issued patent document). The PALM system is the primary workhorse of the patent examination arm of the USPTO. Every office action is assigned a transaction code, a transaction date, and a

[24] PatentsView is a project funded by the USPTO to disambiguate inventor and assignee names and inventor locations. It uses the public version of the USPTO patent grant XML raw data. The latest data version downloaded for this paper is the November 2016 edition: http://www.patentsview.org/download/. Accessed November 2016.

mail to applicant date and ends up recorded in PALM. Because the user interfaces used in the classification and examination of patent applications link directly to PALM, every transaction or touch of the application is recorded in PALM. PALM includes all examiner actions as well as maintenance information and records the internal movement of documents related to an application within the patent office (for example, when an application is assigned to classification, a Technology Center, or an examiner, or when a document is mailed to or received from the applicant).

The Office of the Chief Economist (the research arm of the USPTO) makes available a subset of the transactions in PALM as PatEx (*Pat*ent *Ex*amination Research Dataset). However, the advantage of collecting data internally within the USPTO is that substantial portions of the internal databases are not made publicly available, including the transaction codes necessary for conducting this study, information on pilot programs, and patent applications not yet public or abandoned before the publication date. The latter omission biases the sample of patents included in PatEx because the abandoned applications are underreported (see Graham *et al.*, 2015).

For this paper, I combine IFW information on petitions that provide information on the Green Technology Program with PALM data on patent applications to study the effects of accelerated patenting on firms. The study could not have been conducted without complete access to these internal patent databases, as the patent-related information necessary to construct the firm-level dataset is not publicly available.

Firm Data

Firm-level data are obtained through the National Establishment Time-Series (NETS) dataset, 2014 edition,[25] compiled by Walls and Associates. NETS is the most comprehensive commercially available census of firm locations in the United States, covering over 58 million establishments (an establishment is any business location, including subsidiaries, independent firms, startups, and headquarters). NETS provides linked firm names and locations to Dun and Bradstreet numbers (DUNS) and employment, sales, ownership (public/private), and industry classification data for each DUNS number. The DUNS number is a unique identifier for a firm location. NETS also provides credit ratings for firms in the form of the Dun and Bradstreet Paydex scores (min score and max score within a year). The database, however, does not provide startup funding information. For a detailed overview of NETS, see (Neumark *et al.*, 2007).

A variable of interest in this study is the amount of venture capital funding that startups in the population of interest obtain in the years after filing patents. This information is obtained from CB Insights, a database that provides annual VC funding history and the current status of the firm (active and privately owned, publicly owned, acquired, or no longer active)[26]. A license to CB insights is provided through my university affiliation. The summary statistics are in Table 2.

Table 2: Summary Statistics

	(Obs)	(Mean)	(SD)	(Min)	(Max)
Treated Patents Green (firm-year)	262	2.648	3.564	1	28
Avg Number Patent Claims	843	18.669	7.563	2	62
PaxDexMin (Credit Rating)	864	68.894	12.174	2	80
PaxDexMax (Credit Rating)	864	76	7.514	3	84
Sales (USD)	1461	6.037M	17.6M	6000	157M
Employment	1465	34.76519	103.8765	1	1200
Venture funding (USD)	264	25.2M	44.4M	10000	373M

[25] The NETS license was obtained through the author's collaboration with the United States Patent and Trademark Office.

[26] Other data sources considered for funding information are PREQIN and Thompson ONE, but CB Insights was chosen because it contains more complete historical data for the population of interest.

Disambiguation and Matching Techniques

Firm names can vary over time (for example, Apple Computer became Apple Inc.), and firm names in patent applications can appear with a variety of spellings, abbreviations, and typos. The USPTO does not modify in any way firm names (assignees), nor does it assign a unique identifier to each firm in either its internal or publicly available datasets. As a result, the firm names in the USPTO databases reflect the spellings used in their patent applications. To mitigate the resulting confusion, the USPTO funded an inventor and assignee disambiguation project currently under contract, PatentsView, which assigns a unique identifier to each firm and standardizes the assignee names for all granted patents. A large part of PatentsView's effort was to disambiguate inventor names and locations. For a detailed discussion of the matching method used, discriminative hierarchical co-reference, see Wick *et al.* (2012); for additional efforts in this direction using machine learning, see Kim *et al.* (2016). Due to the higher quality of the assignee data provided by PatentsView, the PatentsView assignee data are merged with the PALM assignee data for our population of interest, using string matching techniques. The resulting standardized firm names are then matched to the NETS database. The matching method is probabilistic record linkage using bigrams (two letter strings), first introduced by Fellegi and Sunter (1969), provided as a Stata package by Blasnik (2010), and improved by Wasi and Flaanen (2015). The contributions of Wasi and Flaanen (2015) enhance the matching accuracy through firm name and address standardization before the processing steps. The results are further manually checked in cases where the match is weak and multiple potential matches are available. Dun and Bradstreet's Hoovers database is used as an additional manual search check step where necessary (i.e., in the case of firms with very similar names across different states).

Final Sample[27]

The sample is US-based startups with patenting activity at the time of the Green Technology Program that identified at least one patent application as green technology. The analysis is performed at the firm-year level. Startups are first identified by filtering assignees of patents in the program based on the fees they paid (startups qualify for small entity fees) and then manually checked against NETS to ensure they are indeed startups. One example of small entities that are not startups are universities (which as non-profits would also qualify for small entity fees); they are excluded from the analysis. Additionally, each individual firm in my main sample (treated and control firms) is checked using Dun and Bradstreet Hoovers to determine if it is a US subsidiary or a foreign multinational; subsidiaries of foreign firms are excluded. The final sample includes 223 startups with patenting activity in the treatment interval that were considered for the Green Technology Program 2009-2012, corresponding to 1,472 firm-years, of which 169 firms were treated (granted accelerated examination).

For the second difference-in-differences analysis, the comparison group was determined using text matching to find the most technologically comparable firms to the treated firms, which yielded an additional 29 startups. The startups were obtained through a run of about 9 billion text comparisons under a 'green tech' classification algorithm introduced in the Methodology section. This part of the analysis required custom written computer code.

In addition to the Green Technology Program firms, data on all green technology industry startups with patenting activity but that did not participate in the Green Technology Program are collected. These 2,013 firms are used to determine a third comparison group for the treated firms as a

[27] Additional details regarding the data will be made available in a working paper in June 2017, after the working paper passes through the customary disclosure processes and peer review at the USPTO.

robustness check, using Coarsened Exact Matching (CEM); the results are in the robustness check section. Additionally, the 3,536 granted patents in the Green Technology Program are compared to all 1,266,609 granted patents filed in the treatment period (2009-2012) to determine whether the coverage of the Green Technology Program's accelerated patenting treatment included all eligible firms. The description of the natural language processing classification is included as a subsection of the robustness checks.

METHODOLOGY

The methods used include standard difference-in-differences, continuous treatment difference-in-differences, CEM, and textual similarity analysis, used for the classification of green technology patents. The treated firms are all firms with at least one patent accepted for acceleration in the Green Technology Program, and the treatment time period is 2009-2012. All models are at the firm-year level. The difference-in-differences specifications include all "green technology" startups with at least one patent accepted for accelerated examination through the Green Technology Program as the treated firms, and all startups that were rejected for acceleration in the Green Technology Program as the control group. To determine whether there is a selection effect (i.e., firms that could have been treated did not get treated), a large-scale text analysis and matching effort is done. A second difference-in-differences analysis was run using the same classification algorithm to identify a comparison group of most-similar-to-treated contemporary startups. Additionally, a CEM analysis is performed to determine *any* green technology sector firms comparable to the treated firms and to determine the average treatment effect. The outcomes of interest regarding startups are sales, employment, and venture funding.

The hypotheses follow the effects of a reduction in patent pendency uncertainty on startup outcomes:

Hypotheses

H1: Startups treated with accelerated patent examination generate higher sales than untreated startups.

H2: Startups treated with accelerated patent examination hire more compared to startups not treated.

H3: Startups treated with accelerated patent examination yield higher venture funding compared to those not treated.

Variables

The dependent variables for the first two hypotheses, pertaining to firm sales and employment, are derived from the business census database NETS. Venture funding data, at the funding per year level, are obtained from CB Insights. Firm patenting characteristics and patent level data are derived from the main two internal USPTO patent databases, IFW and PALM. As patents with more granted claims cover more of their field and are more likely to be valuable, some specifications include a control for the average number of patent claims at the firm-year level. Credit ratings are used as controls and are obtained from NETS.

Difference-in-Differences Base Model and Variables

The three firm outcomes analyzed are firm employment, firm sales, and annual venture funding. First, a standard difference-in-differences approach is used, where $DID_{it} = Post_t * Treated_i$; $Post_t = 0$ for years prior to 2009 (the beginning of treatment) and $Treated_i$ is the treatment dummy (1 for treated firms; 0 otherwise):

$$\ln(Employment_{it}) = \alpha + \gamma \cdot Post_t + \delta \cdot Treated_i + \mu \cdot DID_{it} + X'_{it}\beta + \varepsilon_{it} \quad (1)$$

$$\ln(Sales_{it}) = \alpha + \gamma \cdot Post_t + \delta \cdot Treated_i + \mu \cdot DID_{it} + X'_{it}\beta + \varepsilon_{it} \quad (2)$$

$$\ln(Venture\ Funding_{it}) = \alpha + \gamma \cdot Post_t + \delta \cdot Treated_i + \mu \cdot DID_{it} + X'_{it}\beta + \varepsilon_{it} \quad (3)$$

where *X* is the vector of controls, which depending on the model include $Credit_{it}$ (credit rating is used as a measure of trustworthiness and the financial strength of the firm, using numerical Paydex ratings from Dun and Bradstreet[28]); firm fixed effects; year fixed effects; $Missing\_Credit_{it}$, an indicator of a lack of any credit rating in a year; $Missing\_Credit_{it} * Treated_i$, the interaction between missing a credit rating and the treatment dummy; and the average number of claims in a year per firm, $AvgClaims_{it}$, a measure of the breadth of the firm's patenting (the more claims the firm can produce, the more valuable its IP portfolio will be). Clustering of errors is done at the firm level. All outcome variables are annual.

The NAICS code system is also used for the CEM robustness check and to find the population of green technology firms in NETS (as outlined in the Appendix). The log transformation of the three dependent variables is appropriate here as the distributions of sales, funding, and employment are skewed, for example because many firms get little to no funding (distributions were plotted and are omitted here for brevity).

The following section outlines the robustness checks, including a text analysis and the CEM approach to handle potential selection issues and implementation details. It is possible, for example, that firms eligible for treatment are simply unaware of the program and as such are not

[28] Paydex is a credit rating system for companies produced by Dun and Bradstreet and range from 0 to 100 (best). The values of Paydex reported in NETS are the minimum Paydex in a year (PaydexMin) and the maximum Paydex in a year (PaydexMax). I also include the lack of any credit rating in a year for the firm as a separate control, as lack of credit rating may signify that the firm was not able to gain access to credit that year, which would not be captured by a credit score.

treated; the text analysis addresses this concern and is outlined next and discussed in detail in the section which covers robustness tests.

Natural Language Processing Matching

Eligibility for the Green Technology Program was based exclusively on the topic of the patent application – the patent had to pertain to environmental technology, defined broadly as clean energy generation, recycling, or energy consumption reduction. The decision was based solely on a review of the claims of the patent application and was made by a patent petition adjudicator.[29] Given that *the decision to treat* was made based *on the content of the patent claims text* by a patent official and not based on firm financials or other firm-level characteristics, it is appropriate to determine the population of eligible firms based on textual similarity. To determine the population of all startups eligible for treatment, I implement a classification of all patents in the eligible treatment period (2009-2012) into "green tech" (eligible for treatment) or "not green tech" (not eligible for treatment) using textual similarity and standard text processing methods. The claims are tokenized, stop words and other words common to patents are stripped, and a stemmer is employed. Over 4.46 billion pairwise comparisons are computed to determine the green/not green classification for every patent filed between 2009 and 2012, and the output yields that the vast majority of eligible firms were in fact treated. The same algorithm is run for the three years prior to the treatment period to identify additional firms that are comparable to the treated firms from a technological standpoint.

To determine which patent applications outside the Green Technology Program would have qualified for it, one must look at the language of the patent claims. To ascertain whether green

[29] Private communication, USPTO officer.

technology patents have a different linguistic corpus distribution than the overall patent corpus, a word rank-frequency analysis is conducted. It is well-known that the rank distribution for large corpora is described by Zipf's law, which states a linear relationship in a log-log rank-frequency plot of a natural language (Manning & Schütze, 1999). This is equivalent to saying that the rank distribution has the form $p(r) = \frac{1}{r^{\alpha}}$, $\alpha > 1$, where r is the rank Zipf's Law describes as a linear relationship in a log-log rank-frequency plot of a natural language (Pustejovsky & Stubbs, 2012). The distribution of the green technology claims corpus (see Figure 2) is very different from an ideal natural language distribution, which would appear as a line in this log-log plot (see Brown English Corpus discussion in Pustejovsky & Stubbs, 2012; or refer to the Brown Corpus Zipf plot, an acknowledged representative Corpus of the English language). The curved plot for the claims corpus can be explained through the more frequent use of technical terms than in natural language and the legal terms and phrasings specific to patents which do not appear in normal discourse. Patent texts are more akin to research texts than regular use language, which is relevant in choosing a similarity measure. The Jaccard similarity is used here as it has been found to have the highest purity for research corpora (Huang, 2008) and has been found more effective than cosine for online text as well (Leydesdorff, 2008), and because interviews with the USPTO examiners suggest that a vocabulary match is the appropriate comparison. This similarity choice also enables the running of 4.46 billion text comparisons in under a day on a high-end consumer-grade system, with programming implementation choices that optimize execution speed.[30]

[30] The Jaccard similarity-based text comparisons of patent claims texts completed 4,461,000,000 comparisons in about 17 hours on a core i-7 7700K system at 4.85 GHz with 64GB RAM.

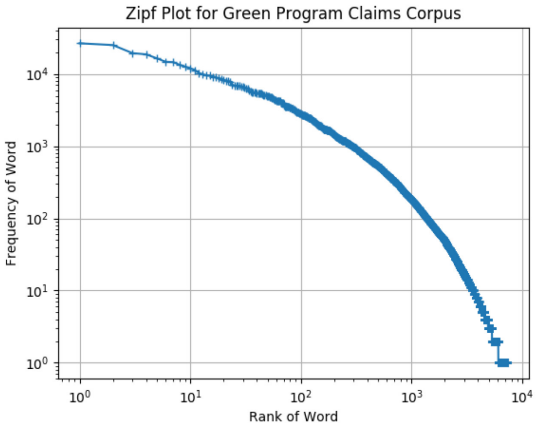

Figure 2. Corpus rank frequency plot of green patent claims in the target treatment period.

The vocabulary of the Green Technology Program patents is vastly different from that of regular patents in the same time period (2009-2012), as shown in the Appendix Table A1. The green corpus showing several easily recognizable green technology terms, such as "power," "light," "wind," "cell," "turbine," and "solar" at the top of the distribution. The words in the table are obtained using a lemmatization process. The WordNet lemmatizer is a standard algorithm to reduce all inflections of words to their stem, that is, the root of the word, or dictionary form. This is important in text analysis as the frequencies of words will not be counted correctly if we do not map the inflections of a word (tense, plurals, etc.) to a single dictionary entry. The WordNet lemmatizer relies on an internal dictionary that maps words to Parts of Speech (POS) tags. Each input word is verified against the WordNet dictionary, a process that is computationally expensive. A widely accepted and much faster approach is to stem words using rules that do not require a dictionary search; the English language has multiple standard *stemmer* implementations that use rules rather than dictionaries to reduce words to their roots. The standard stemmers for English include the Porter and the newer Porter 2 stemmers. For a comprehensive overview of stemmers and lemmatizers, see Manning and Schütze (1999).

The typical patent claims in the green corpus have a fairly limited vocabulary (mean 64 words; standard deviation 25 words; the largest vocabulary is 335 words) after removal of common words and the appropriate reduction of words to their roots using the WordNet lemmatizer. With these descriptive analyses of the green technology patents, I show that the green patents corpus is quite distinct linguistically. However, this does not yet help us determine whether the treatment covered all or most eligible patents. Instead, a more extensive algorithm needs to be introduced.

To determine which patents could have qualified for the Green Technology Program, I apply the following algorithmic steps:

1. Tokenize claims and remove stop words, including patent-specific stop words as defined by the USPTO[31] as well as stop words found by Lexis Nexis.[32]
2. Stem the words using the Porter 2 standard stemmer.
3. Generate the vocabulary set of each patent application.
4. Compute the Jaccard similarity of pairs of (Green Technology Program patent, other patent in 2009-2012) and *discard all the pairs with a similarity below 0.5*[33].
5. Match all patents found in step 4 to the business census data from NETS and discard all firms that were not US-based startups at the time of the program (2009-2012).[34]

The Jaccard similarity is defined as the cardinal of the intersection of two sets over the cardinal of their union, in this case exemplified by two sets of words W_1 and W_2:

[31] USPTO stop words list, http://patft.uspto.gov/netahtml/PTO/help/stopword.htm. Accessed August 30th, 2017.
[32] Lexis Nexis list of patent-specific stop words, http://help.lexisnexis.com/tabula-rasa/totalpatent/noisewords_ref-reference?lbu=US&locale=en_US&audience=online. Accessed August 30th 2017
[33] A threshold of 0.5 is widely accepted as more similar than not for this similarity.
[34] This step is an involved one, as detailed in the earlier "Disambiguation and Matching Techniques" section.

$$Jaccard(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$$

A Jaccard similarity value of above *0.5* indicates that the intersection of the two sets covers at least half of all the elements in the two sets.

A first purpose of using textual analysis is to find *the entire population of firms eligible* for treatment. The 4.46 billion text comparisons corresponding to all green tech treated patents compared to all patents filed in 2009-2012 are obtained from a run in a custom Python script. The output of this process is finding the *patents most similar to the treated patents*, i.e., the patents most similar to the green technology patents. These most similar patents would have been eligible for treatment and can be matched to the firm census to find startups that could have been eligible for treatment. Out of the treatment period, 2009-2012, *only nine additional startups that could have been treated are identified, which implies near-perfect coverage of intended treatment*. Therefore, the firms eligible for treatment per the USPTO guidelines for the Green Technology Program are virtually all included in the treatment group. The natural language processing analysis can be applied for more than determining treatment selection, however.

The first set of Results utilize as a control group the group of firms identified from the USPTO files as 'considered for treatment but not treated' and who are US based green technology startups. I recognize the limitations of this approach, namely that the treatment is not truly randomly assigned, having been decided by a human reader in the USPTO office and not by a random process. While the first analysis does show an effect of treatment, an additional analysis is necessary to determine the *most similar group of startups* to the treated firms which could have been treated but were not. This is a repeat of the algorithm outlined above, but run for three additional years preceding the treatment, and, combined with the nine firms who were arbitrarily

not included in the treatment group during the treatment period, comprises an additional control group, of most technologically similar startups to the treated. This second group is the result of about 9 billion text comparisons and a lookup of the NETS database. The results of both difference-in-differences analyses are consistent.

Implementation Notes

The Natural Language Processing component is custom-implemented in Python,[35] linked to a CSV version of the patent claims dataset of all US patent applications filed 2009-2012, and optimized to run on consumer available hardware. Data analysis and all regressions are done in Stata. Pendency statistics are done in Fluxicon Disco,[36] a process mining commercial software applied to the public transactions database of the USPTO.

EMPIRICAL RESULTS

Analysis Based on Program Treatment and Control

The results in this section use treatment and control as the US startups considered for treatment and given treatment versus the US startups considered under the program and denied treatment. This setup is directly from the setup of the program at the USPTO and is a first analysis; the drawbacks of this setup are explained in the prior section and led to the development of a text-matched comparison group, analysis which follows this section.

Table 3 shows that for a variety of specifications, there is an increase in sales with each treated patent, with the value of treatment for the firm on the order of 30% in extra sales. Regressions using logged outcomes show that the effects of the treatment on employment are higher by over

[35] The Python code for the NLP processing will be made available as part of an online appendix.

[36] I thank Dr. Anne Rozinat and Fluxicon Netherlands, developer of Disco, the process mining software, for graciously waiving licensing fees for the purpose of this academic research.

25% post treatment, as shown in Table 4. These results hold after two years as well (treatment continues to have an effect two years after it occurs). Of the control variables, Missing Credit, an indicator for lacking a credit rating, is found to be significant and detrimental to sales (as would be expected, as it is a negative signal to those outside the firm); the average number of claims made by a firm in its patents in a given year is not significant. It is expected that missing a credit rating will hurt a firm's sales, as lacking a credit rating indicates that the firm has limited borrowing capabilities, which in turn can affect production. Controlling for industry subsectors does not change the effects. The effects of treatment are significant in a variety of specifications.

Table 3: Logged Firm Employment Regressions - Difference-in-Differences Specifications

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.3219***	0.3958 ***	0.3392***	0.4894***	0.2541***	0.33495**
	(0.0561)	(0.0848)	(0.0920)	(0.1174)	(0.0566)	(0.1199)
POST	0.4143***	0.3236	0.2661	0.1954	0.3388***	0.2622
	(0.0927)	(0.1532)	(0.1526)	(0.2100)	(0.0908)	(0.1657)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
NAICS FE					Yes	Yes
Avg Num Clms		0.0047	0.0054	0.0107		0.0049
		(0.0050)	(0.0049)	(0.0073)		(0.0053)
Missing Credit			-0.2698			-0.1457
			(0.1028)			(0.1561)
Missing CreditXTreat						-0.1907
						(0.1481)
PaydexMin Credit				-0.0008		
				(0.0072)		
PaydexMax Credit				-0.0125		
				(0.0170)		
N (firm-years)	1,465	753	753	498	1,465	749
R^2	0.8589	0.8819	0.8848	0.8602	0.8934	0.9333

Robust Standard Errors in parentheses (firm level)
Model (1) is the base model; Model (2) includes average number of claims per patent at the firm year level (a measure of firm IP sophistication). Model (3) includes an indicator for measures of credit rating. There is much less information available for the Dun&Bradstreet PayDex vars in (4) compared to (1) and (2), however missing credit rating may be in itself information, as a signal of firm quality. To test this, missing credit is a separate variable in Model (3). Model (4) includes the numerical credit ratings from Dun&Bradstreet.; models (5), (6) have both NAICS industry and firm fixed effects. Model (6) also shows the interaction between treatment amount and missing credit rating. In all specifications DID remains significant and positive.

Table 4: Logged Firm Sales Regressions - Difference-in-Differences Specifications

	(1)	(2)	(3)	(4)	(5)	(6)
DID	0.3445***	0.3851 ***	0.3180**	0.4714**	0.2835***	0.3694**
	(0.0694)	(0.1058)	(0.1126)	(0.1499)	(0.0732)	(0.1377)
POST	0.4926***	0.4622**	0.3922*	0.3770	0.3885***	0.3530*
	(0.1142)	(0.1630)	(0.1526)	(0.2218)	(0.0960)	(0.1558)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
NAICS FE					Yes	Yes
Avg Num Clms		0.0062	0.0071	0.0138		0.0065
		(0.0058)	(0.0057)	(0.0081)		(0.0062)
Missing Credit			-0.3258**			-0.1489
			(0.1179)			(0.1650)
Missing CreditXTreat						-0.2656
						(0.1549)
PaydexMin Credit				-0.0206		
				(0.0166)		
PaydexMax Credit				-0.0044		
				(0.0096)		
N (firm-years)	1,461	754	754	498	1,450	749
R^2	0.8954	0.9169	0.9192	0.8964	0.9232	0.9333

Robust Standard Errors in parentheses (firm level)
Model (1) is the base model; Model (2) includes average number of claims per patent at the firm year level (a measure of firm IP sophistication). Model (3) includes an indicator for measures of credit rating. There is much less information available for the Dun&Bradstreet PayDex vars in (4) compared to (1) and (2), however missing credit rating may be in itself information, as a signal of firm quality. To test this, missing credit is a separate variable in Model (3). Model (4) includes the numerical credit ratings from Dun&Bradstreet.; models (5), (6) have both NAICS industry and firm fixed effects. Model (6) also shows the interaction between treatment amount and missing credit rating. In all specifications DID remains significant and positive.

A difference-in-differences regression with annual venture capital funding as the outcome variable is also run and yields that the difference in funding between firms treated with the accelerated patenting versus firms not treated is found in the base model to be on the order of 58% . This is a large effect, the largest of the three firm outcomes, and is supported by theory - investment is delayed when there is uncertainty; with a decrease in uncertainty in terms of IP assets of the startups, due to the treatment in the Green technology program, investment should increase. The data is however much more limited, and I was not able to run the model with as many specifications

as the other two outcomes. In addition to the base model, the model involving credit rating shows that not having a credit rating is detrimental to venture funding as well, similar to the results obtained for the sales outcome. The venture funding regressions are in Table 5. These results show that in a reduction of uncertainty in IP outcomes, additional investment funding is made available, and firms who benefit from this reduction of uncertainty fare better. However, I recognize that the analysis in this section is impaired by the fact that the assignment of patent applications to treatment and control was not, as is often the case with programs run within very specific legal limits, truly random. Assignment is based on whether the patent application is "truly green technology," where some patents may be clearly green versus others only partly green, in the words of one of the government employees interviewed (5/1/17), and not a randomized experiment[37]. This limitation is in the setup of the program. In order to address it, I proposed the difference-in-differences approach with a text-matched control group, method described earlier. I also ran Coarsened Exact Matching as a separate check with a well known method, which produced yet another comparison group. The hypotheses were validated by these two additional methods. Future iterations of the paper will focus on the text matched and CEM approaches.

[37] From the same interview, it was suggested that often the legal framework specific to the agency may prevent a truly randomized experiment affecting actors outside the agency. This may vary from one government agency to another, but in the question of uncertainty reduction in IP, this may be the best setup to test, given existing policy changes and an analysis of internal documents.

Table 5: Logged Annual Venture Funding Regressions (Preliminary)

	(1)	(2)	(3)	(4)
DID	0.5844*	0.4484^{+}	0.3984	0.3789
	(0.0374)	(0.2581)	(0.2457)	(0.3550)
POST	0.0373	-0.1968	0.1352	0.2630
	(0.2952)	(0.3071)	(0.2984)	(0.4056)
Firm FE	Yes	Yes	Yes	Yes
NAICS FE	No	No	No	Yes
Missing Credit		-0.7119*		0.0206
		(0.2747)	(0.4811)	
Avg Num Clms			-0.0072	-0.01251
			(0.0130)	(0.0138)
_cons	13.7958***	14.1660***	14.2043***	17.1856***
	(0.2483)	(0.2847)	(0.4065)	(0.4495)
N (firm-years)	264	264	175	166
R^2	0.7491	0.7607	0.7743	0.7673

Robust Standard Errors in parentheses (firm level)

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Model (1) is the base model; Model (2) includes an indicator for measures of credit rating. Model (3) includes average number of claims per patent at the firm year level (a measure of firm IP sophistication). Model (4) includes NAICS level fixed effects. There is far less data available on venture funding than other outcomes.

A separate model is that of a continuous treatment (where firms with over one patent in the green technology program receive a higher level of treatment) as the construction in (Acemoglu & Finkelstein, 2008), which shows that the added benefit of one additional treated patent for the average firm is of the order of over $200k in sales. This setup is shown in Table 6.

Table 6: Firm sales regressions diff-n-diff results - continuous treatment

	(1)	(2)	(3)	(4)	(5)	(6)
DID	476546.3***	374939.2 **	396495.5**	506349.2*	356174.6*	175248.8*
	(129497.2)	(121674.7)	(123635.6)	(205605.4)	(154181.2)	(74531.82)
Firm FE	Yes	Yes	Yes	No	Yes	Yes
Year FE	No	No	No	Yes	Yes	Yes
Avg Num Clms		101614.2^{+}	189701.6			85811.21
		(60648.51)	(130548.5)			(67263.57)
Missing Credit					860148.8	-717067.3
					(723675.1)	(1583007)
MsngCrdXTrtm						100553.7
						(335091.7)
Paydexmin			-25394.05			
			(52824.19)			
Paydexmax			-146449.3			
			(200256.6)			
N (firm-years)	1,461	754	498	1,461	1,461	754
R^2	0.8526	0.8467	0.8445	0.0647	0.8721	0.9102

Robust Standard Errors in parentheses ((1)-(5) firm level clustering; (6) year level)
Model (1) is the base model; Model (2) includes average number of claims per patent at the firm year level (a measure of firm IP sophistication). Model (3) includes measures of credit rating from Dun&Bradstreet. There is much less information available for these Dun&Bradstreet PayDex vars in (3) compared to (1) and (2), however missing credit rating may be in itself information, as a signal of firm quality. To test this, missing credit is a separate variable in Model (5). Model (4) is simply a model with only year fixed effects for comparison; models (5), (6) have both firm and year fixed effects. Model (6) also shows the interaction between treatment amount and missing credit rating. In all specifications DIDD remains significant and positive.

Analysis Based on Text Matched Treatment and Control

Maintaining the treatment group decided by the USPTO, an additional comparison group is obtained through the text matching algorithm outlined in the methods section. The text matched group is the most similar technologically to the treated group; firm characteristics from NETS are used to run a pre-trends check and both employment and sales pre-trends are parallel (see Figure 3 below). The firms found through the text matching group would have been eligible for treatment but were not treated. All are US based startups.

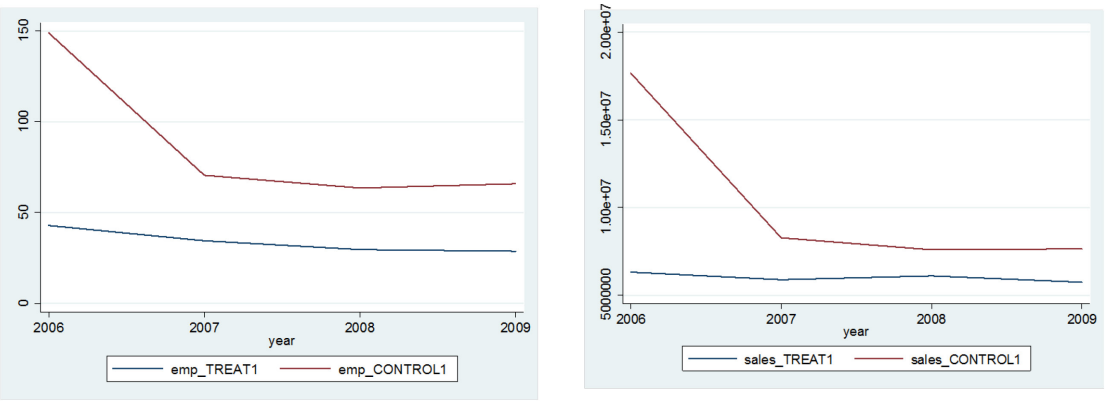

Figure 3. Pre-trends (treatment begins in year 2009) for the NLP matched control group.

The results in this case are consistent for both employment and sales outcomes with the prior section, and show double digit percentage increases for treated startups, as shown in Table 7 and Table 8:

Table 7: Logged Firm Employment Regressions - NLP Matched - Difference-in-Differences Specifications

	(1)	(2)	(3)	(4)
DID	0.4008*	0.4477 *	0.4177*	0.3792^{+}
	(0.1610)	(0.2122)	(0.2084)	(0.2197)
POST	0.2534*			
	(0.1254)			
Firm FE	Yes	Yes	Yes	Yes
Year FE		Yes	Yes	Yes
Missing Credit			-0.1278^{+}	
			(0.0697)	
Missing CreditXTreat				-0.1298 $^{+}$
				(0.0721)
N (firm-years)	1,571	1,571	1,571	1,571
R^2	0.8615	0.8736	0.8743	0.8602

Table 8: Logged Firm Sales Regressions - Difference-in-Differences NLP Matched

	(1)	(2)	(3)	(4)
DID	0.5563**	0.6640 **	0.6402*	0.6090*
	(0.1973)	(0.2346)	(0.2237)	(0.2371)
POST	0.1930			
	(0.1592)			
Firm FE	Yes	Yes	Yes	Yes
Year FE		Yes	Yes	Yes
Missing Credit			-0.1014	
			(0.0700)	
Missing CreditXTreat				-0.1043
				(0.0718)
N (firm-years)	1,570	1,570	1,570	1,570
R^2	0.8947	0.9033	0.9036	0.9036

The text matching technique for constructing a control group is applicable to other fields beyond patents, for example when finding comparable firms based on a finer classification than industry level (for instance, using text of the product descriptions), for grant applications, or whenever the criteria for treatment may be extracted from a body of text. Patents certainly hold an advantage in that the texts all follow the same structure and specialized language, and roughly similar vocabulary size, whereas other bodies of texts may require slightly different analyses, perhaps changes in the similarity measure or other alterations. Despite the limitations, the method of constructing a text matched control group will likely have many future applications.

ADDITIONAL ROBUSTNESS CHECKS

To determine whether selection into the program plays a role in the results, I run Coarsened Exact Matching (CEM) to find comparable firms in the population of all green technology firms with patenting activity.

All of the identifiable characteristics of the treated firms are used to create a comparison group of green technology firms with patents that did not obtain an accelerated patent. To implement this approach, the *entire population of green technology firms in the US in the treatment period* is determined by using the largest available commercial business census, the NETS database. The approach is as follows:

1. Determine all NAICS codes pertinent to green technology (using the Bureau of Labor Statistics 6 digit NAICS code classification of green technology industry sectors, a much finer sub-industry classification that often yields the family of products, such as thermal control valves, hydro power, biofuels, etc.).
2. Determine the population of all green technology firms in the treatment period (2009-2012) by taking a subset of the NETS business census with the NAICS codes from step 1 and the firms active in the treatment period (2009-2010).
3. Match the firm names from step 2 with the complete population of patents in the US to determine which firms from the population of green technology firms had patenting activity.
4. Match treated firms using the set in step 3 to determine a comparison group based on firm and patenting characteristics (pre-treatment size, pre-treatment sales, NAICS industry classification pre-treatment, geographic region, and creditworthiness).
5. Estimate the average treatment effect on the treated by using the matched group from 4.

The matching method used is the CEM approach (Blackwell *et al*., 2009; Iacus *et al*., 2011; Iacus *et al*., 2012), which has been shown to work well with both continuous and discrete characteristics. No change in the direction or magnitude of the treatment is found between firms that were *not treated but are comparable to treated firms* as obtained through CEM (as shown in Table 9). With

a variety of matching criteria, the average treatment effect on sales over the treated (covering the treatment period) remains in the millions of dollars and is highly significant. This serves as an additional robustness check and strengthens the main result.

Table 9: Coarsened Exact Matching Results - Robustness Check for Sales

	SATT	(SE)	N Treated Matched
Match pre-treat sales empl	4884065***	(1025397)	82
Pre-treat sales, empl, PayDexMin	5124341***	(685616)	81
Pre-treat sales, empl, PayDexMax	5112332***	(700833)	81
Pre-treat sales, empl, NAICS	3697812**	(1147451)	55
Pre-treat sales, empl, PayDex(Min&Max),NAICS	4481249***	(626767.7)	54
Pre-treat sales, empl, NAICS, and Region	4386675***	(717543.5)	53
Pre-treat sales, empl, PayDex(Min&Max), SIC2, Region	3689133***	(771428)	55

Further Checks

In addition to the text analysis and CEM approaches, I am working on additional robustness checks, including matching on use of a lawyer pre-treatment, which may indicate higher firm sophistication, as well as fee data (if the firm considers itself a "small entity," for example). Further, I plan to use the matched groups found through CEM in additional robustness checks and am exploring the possibility to use the text analysis as a gradient of green technology, as to enable a regression discontinuity design where firms that had very close, but not close enough technology to qualify for treatment are compared to the treated firms. While formal grading criteria for what constituted "green technology" were not found in the program studied here, further interviews will be attempted at the USPTO in order to inform further text algorithm changes in view of a scale of "green technology" that would better emulate the decisions of the patent examiner.

CONCLUSIONS

Firms that accelerate their patents fare much better in terms of firm outcomes than firms that do not accelerate their patents and have an unknown and highly variable patent pendency. This paper presents a unique setup that allows for studying the effects of a reduction in patent pendency on

firm outcomes, namely a government program targeted at encouraging innovation in green technology by reducing the patent pendency time. The text similarity approach used to determine the population of all eligible patents for acceptance into the Green Technology Program is novel for the management innovation literature, has wide applications beyond this paper, and may be useful for determining the population of eligible applicants and potential selection issues for any other programs with application materials that can be textually compared.

The findings in this study have implications for firm strategy, showing there is a high value in accelerating patents and that firms should take advantage of any opportunities under their control to reduce patent pendency, especially if they need the IP for relationships with VCs, manufacturing partners, or customers. Further, this paper has some policy implications, showing that governments could successfully use IP levers involving patent examination timing to increase investments in targeted industry sectors. Overall, the contributions of this paper are three-fold: a test of the effects of a reduction of uncertainty in patenting on firms based on a new policy, the construction of a new dataset from the government source and insights based on internal research, and introduction of a text matching method for use in difference-in-differences analyses.

REFERENCES

Acemoglu, D., & Finkelstein, A. 2008. Input and technology choices in regulated industries: Evidence from the health care sector. ***Journal of Political Economy,*** 116(5): 837-880.

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In ***The rate and direction of inventive activity: Economic and social factors***, pp. 609-626. Princeton, NJ: Princeton University Press.

Anton, J. J., & Yao, D. A. 1994. Expropriation and inventions: Appropriable rents in the absence of property rights. ***The American Economic Review,*** 84(1): 190-209.

Anton, J. J., & Yao, D. A. 2004. Little patents and big secrets: Managing intellectual property. ***The RAND Journal of Economics,*** 35(1): 1–22. Available at www.jstor.org/stable/1593727.

Arora, A., & Gambardella, A. 2010. The market for technology. ***Handbook of the Economics of Innovation,*** 1: 641-678.

Athey, S., & Imbens, G. W. 2015. Machine learning methods for estimating heterogeneous causal effects. a***rXiv Stat.ML,*** 1050: 5.

Blackwell, M., Iacus, S. M., King, G. & Porro, G., 2009. cem: Coarsened exact matching in Stata. ***Stata Journal,*** 9(4): 524-546.

Blasnik, M. 2010. RECLINK: Stata module to probabilistically match records. ***Statistical Software Components***.

Fellegi, I. P., & Sunter, A. B. 1969. A theory for record linkage. ***Journal of the American Statistical Association,*** 64(328): 1183-1210.

Frakes, M. D., & Wasserman, M. F. 2017. Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from micro-level application data. ***Review of Economics and Statistics,*** 99(3): 550-563.

Gans, J. S., Hsu, D. H., & Stern, S. 2008. The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays. ***Management Science,*** 54(5): 982-997.

Galasso, A., & Schankerman, M. A. 2015. Patent rights, innovation and firm exit. ***NBER Working Paper*** No. 21769.

Graham, S. J. H, Marco, A. C., & Miller, R. 2016. The USPTO patent examination research dataset: A window on the process of patent examination. ***Georgia Tech Scheller College of Business Research Paper*** *No. WP* 43.

Farre-Mensa, J., Hegde, D., & Ljungqvist, A. 2016. The bright side of patents. ***NBER Working Paper***.

Hegde, D., & Luo, H. 2018. Patent publication and the market for ideas. ***Management Science***, 64(2): 652–672

Huang, A. 2008. Similarity measures for text document clustering. ***Proceedings of the Sixth New Zealand Computer Science Research Student Conference 2008***. Christchurch, New Zealand, pp. 49-57.

Iacus, S. M., King, G. & Porro, G. 2011. Multivariate matching methods that are monotonic imbalance bounding. ***Journal of the American Statistical Association,*** 106(493): 345-361.

Iacus, S. M., King, G. & Porro, G. 2012. Causal inference without balance checking: Coarsened exact matching. ***Political Analysis,*** 20(1): 1-24.

Kim, K., Khabsa, M., & Giles, C. L. Random Forest DBSCAN for USPTO inventor name disambiguation. 2016. ***arXiv preprint:*** *1602.01792.* Available at https://arxiv.org/pdf/1602.01792.pdf.

Lerner, J. Patenting in the shadow of competitors. 1995. ***The Journal of Law and Economics,*** 38(2): 463-495.

Lerner, J. 1996. The government as venture capitalist: The long-run effects of the SBIR program. ***NBER Working Paper*** No. 5753.

Marco, A. C., Sarnoff, J. D., & deGrazia, C. 2016. Patent claims and patent scope. ***USPTO Economic Working Paper***. Available at http://dx.doi.org/10.2139/ssrn.2844964.

Manning, C. D., & Schütze, H. 1999. ***Foundations of statistical natural language processing***. Cambridge, MA: MIT Press.

Neumark, D., Zhang, J., & Wall, B., 2007. Employment dynamics and business relocation: New evidence from the National Establishment Time Series. In ***Aspects of worker well-being***, pp. 39-83. Emerald Group Publishing Limited.

Pustejovsky, J., & Stubbs, A. 2012. ***Natural language annotation for machine learning: A guide to corpus-building for applications***. Sebastopol, CA: O'Reilly Media.

Sampat, B., & Williams, H. L. 2015. How do patents affect follow-on innovation? Evidence from the human genome. ***NBER Working Paper*** No. 21666.

Sidak, J. G. & Skog, J. O. 2015. Attack of the shorting bass: Does the inter partes review process enable petitioners to earn abnormal returns. ***UCLA L. Rev. Discourse,*** 63(120).

Sukhatme, N. U., & Cramer, J. N. L. 2014. Who cares about patent term? Cross-industry differences in term sensitivity. Working Paper.

Toole, A. A., & Turvey, C. 2009. How does initial public financing influence private incentives for follow-on investment in early-stage technologies? ***The Journal of Technology Transfer,*** 34(1): 43-58.

Wasi, N., & Flaaen, A. 2015. Record linkage using Stata: Preprocessing, linking, and reviewing utilities. ***Stata Journal,*** 15(3): 672-697.

Wick, M., Singh, S., & McCallum, A. A discriminative hierarchical model for fast coreference at large scale. ***Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1***. Association for Computational Linguistics, 2012.

APPENDIX – Chapter 3

A comparison of the top of the green technology claims vocabulary distribution and that of regular patents is shown below:

Table A.1. Comparison of green technology patents corpus and all patent applications, corpus 2009-2012 (top of distribution); stop words are removed and the NLTK WordNet Lemmatizer is used. Notice the very different distributions, with the green corpus showing several easily recognizable green technology terms, such as "power," "light," "wind," "cell," "turbine," and "solar":

Top Words Green Patent Claims Corpus 2009-2012	**Top Words Patent Application Claims Corpus 2009-2012**
method	method
system	device
layer	system
power	data
device	include
light	portion
plurality	plurality
wind	signal
cell	unit
include	layer
turbine	surface
control	form
surface	configure
portion	information
configure	base
material	control
energy	apparatus
solar	image
couple	receive
form	group