UC Berkeley
UC Berkeley Electronic Theses and Dissertations

**Title**
Measurement and Causal Inference in Patent Strategy

**Permalink**
https://escholarship.org/uc/item/0791g94s

**Author**
Kuhn, Jeffrey Michael

**Publication Date**
2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

# Measurement and Causal Inference in Patent Strategy

By

Jeffrey M. Kuhn

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rui de Figueiredo, Chair
Professor Noam Yuchtman
Professor Brian Wright
Professor Edward Augenblick
Professor Robert Merges

Fall 2017

# Measurement and Causal Inference in Patent Strategy

Copyright 2017

by

Jeffrey M. Kuhn

# Abstract

Measurement and Causal Inference in Patent Strategy

by

Jeffrey M. Kuhn

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor Rui de Figueiredo, Chair

Determining the effect of patents on firms has long presented significant challenges to innovation scholars due to the specific and idiosyncratic nature of patent protection. This thesis develops new techniques for measuring the impact, technological relatedness, and extent of patents. It then describes two new approaches to performing causal inference on patent protection. Finally, it applies these new measures and causal instruments to investigate the effect of receiving a broader patent on patent sales and the effect of winning a patent race on follow-on innovation.

Patent scope is central to the sale of ideas, which can spur economic growth and provide significant gains from trade. Awarding an inventor a patent on a new idea partially solves a commitment problem that would otherwise prevent the inventor from selling the idea. (Arrow, 1962). In the absence of a patent, a prospective buyer cannot credibly promise not to steal the idea should the inventor reveal it, while the inventor cannot credibly promise to reveal the idea should the prospective buyer pay for it. A firm's ability to use a particular patent to overcome this transactional hurdle derives from two factors: (1) the scope of the patent's legal right to exclude and (2) the effectiveness of that legal right in providing market exclusivity. This thesis first shows that a broader patent is more likely to be sold by employing a causal instrument that provides a plausibly exogenous shock to the scope of a patent's legal right to exclude, holding fixed the underlying idea. It then examines variation in the effectiveness of the right by interacting the instrument with endogenous firm, industry, and market characteristics. These results shed light on how firms profit from innovation and also connect the important but understudied market for patents, widely believed to be illiquid and inefficient, with fundamental research about how markets function in other contexts.

Competition between firms to invent and patent an idea, or "patent racing," has been much discussed in theory, but seldom analyzed empirically. This thesis introduces an empirical way to identify patent races, and provides the first broad-based view of them in the real world. It reveals that patent races are common, particularly in information-technology fields. The analysis is then extended to identify the causal impact of winning a patent race, using a regression-discontinuity approach. It shows that both the winners and losers of patent races typically receive patent protection, but that the winners receive much broader patent scope. It also shows that patent race winners do significantly more follow-on innovation, and the follow-on research that they do is more similar to what was covered by the patent.

Underlying both of these analysis is a new measure of patent-to-patent similarity. Current measures of patent similarity rely on the manual classification of patents into taxonomies. This thesis describes and validates a machine-automated measure of patent-to-patent similarity developed by leveraging information retrieval theory and Big Data methods. It also demonstrates that the measure significantly improves upon existing patent classification systems.

## Dedication

To my wife Dana, without whose support this thesis would not have been possible.

## Acknowledgments

I would like to thank my parents, Mike and Darlene, for their consistent love and support during my graduate studies and entire life. I would also like to thank Adam Lugowski, Stephen Glade, Nirav Thakor, Santiago Truffa, and Yoonha Kim for enriching my life. I would also like to thank my co-authors, Kenneth Younge, Neil Thompson, and Alan Marco. I would like to especially thank Rui de Figuieredo, Noam Yuchtman, Robert Merges, Brian Wright, Ned Augenblick, and the other fantastic faculty members at the University of California.

**Contents**


**1 Introduction** **1**

**2 Patent-to-patent Similarity: A Vector Space Model** **4**
2.1 Introduction . . . 4
2.2 The Vector Space Model . . . 5
2.3 Validation . . . 9
2.4 Classification . . . 16
2.5 Conclusion . . . 19

**3 Does Winning a Patent Race Lead to More Follow-on Innovation?** **21**
3.1 Introduction . . . 21
3.2 Patent Races in Theory . . . 22
3.3 Patent races in the wild . . . 23
3.3.1 Continuous Technology . . . 23
3.3.2 Winning a Patent Race . . . 24
3.3.3 How the patent system lets us observe patent races . . . 24
3.3.4 Patent priority . . . 26
3.3.5 Data . . . 26
3.3.6 Defining Patent Races . . . 28
3.4 Where do Patent Races Happen? Who races? . . . 31
3.4.1 Description of Patent Races . . . 31
3.4.2 Technology Areas . . . 32
3.4.3 Firms . . . 33
3.5 Methodology for Obtaining a Causal Estimate of the Effect of Winning a Patent Race . . . 35
3.5.1 Regression Discontinuity Approach . . . 35
3.5.2 Testing Whether those that Win and Lose Patent Races Look Similar, Pre-treatment . . . 36
3.5.3 Firm Portfolio Approaches to Patent Racing . . . 39
3.6 Causal Effect of Winning a Patent Race . . . 40
3.6.1 Patent Protection Obtained and Maintained . . . 41
3.6.2 Effect on Firm Level Outcomes . . . 42
3.6.3 Follow-on Research, Conditional on the Issuance of Both Patents . . . 43
3.7 Conclusion . . . 46

**4 Property Rights and Frictions in the Sale of Patents** **48**
4.1 Introduction . . . 48
4.2 The Importance of Patent Ownership . . . 50
4.3 Institutions and Hypotheses . . . 52
4.3.1 The Delineation of the Right to Exclude . . . 52
4.3.2 The Effectiveness of the Right to Exclude . . . 54
4.4 Methodology . . . 56

4.4.1 Sample Selection and Data . . . . . 57
4.4.2 Identification Strategy . . . . . 61
4.4.3 Descriptive Statistics . . . . . 62
4.4.4 Econometric Specification . . . . . 65
4.5 Results . . . . . 67
4.5.1 Main Results . . . . . 67
4.5.2 Industry and Market Characteristics . . . . . 71
4.6 Conclusion . . . . . 78

**5 Conclusion** **81**

List of Tables


1 Summary statistics. . . . 7
2 Predictions of patent citation. . . . 15
3 Predictions of patent classification. . . . 16
4 Summary statistics. . . . 27
5 Comparison of the firm characteristics of twin patents to firm characteristics of all patents. Full sample 2005. . . . 34
6 Comparison of the firm characteristics of twin patents to firm characteristics of all patents, on a per-employee basis . . . . 35
7 Comparison of the (logged) firm characteristics for twin patents to firm characteristics for all patents. Full sample 2005. . . . 35
8 Application-level balance. . . . 36
9 Financial characteristics (logged). . . . 38
10 Firm-level patent characteristics. Covariates logged. . . . 38
11 Patent characteristics. Covariates logged. . . . 38
12 Patent protection and maintenance results . . . . 41
13 Firm-level financial results. . . . 43
14 Patent-level citation results. . . . 44
15 Variable definitions and summary statistics. . . . 58
16 Summary statistics by industry . . . . 60
17 First stage by industry. . . . 61
18 Balance statistics (t-tests). . . . 62
19 Sales and forward citations. . . . 64
20 Sales and technological relatedness. . . . 65
21 Main results by industry, endogenous regression. . . . 67
22 Main results by industry, reduced form. . . . 68
23 Main results by industry, instrumental variable. . . . 69
24 Instrumental variable results by industry with firm size interactions. 70
25 Main results by industry, logistic regression. . . . 70
26 Logistic regressions by transfer type and industry-reported patent effectiveness. . . . 72
27 Logistic regressions by transfer type and industry-reported patent effectiveness. . . . 74
28 Logistic regressions by transfer type and market characteristics. . . . 75
A1 Instrumental variable regressions by CNW-reported patent effectiveness overall and for exclusion. . . . 90
A2 Instrumental variable regressions by CNW-reported patent effectiveness for cross-licensing and negotiation. . . . 91
A3 Instrumental variable regressions by transfer type and market characteristics, clustered standard errors. . . . 92
B1 Correlation Matrix. . . . 95
B2 Summary statistics by CNW industry. . . . 96

**List of Figures**


1 **Angular separation in a vector space model.** . . . 8
2 **Distribution of similarity by sample.** . . . 8
3 **Patent similarity within firms.** . . . 10
4 **Patent similarity within patent families.** . . . 11
5 **Probability that a patent is sold as a function of market characteristics, by examiner score, predicted based on logistic regression estimates.** . . . 13
6 **Patent similarity within patent classifications.** . . . 17
7 **Overlapping patent classifications.** . . . 18
8 **Histogram of rejections by difference in filing date.** . . . 29
9 **A continuation of Figure 8, with a detailed view of years 0-2.** . . . 29
10 **Similarity (as measured by cosine distance) between patents in the same (main) patent class or patent twins. Scores only calculated for cases with similarity greater than a (low) minimum threshold for computational reasons.** . . . 30
11 **Percentage of patents narrowed during examination at the USPTO.** 31
12 **Patents in patent races.** . . . 32
13 **The percentage of patents issued in 2005 that are in at least one twin, binned by USPTO technology center. The networking technology center is omitted because it was started in 2005 and not yet issuing patents.** . . . 32
14 **Patents in races, as a percentage of patents with novelty rejections (2005).** . . . 33
15 **Sample Composition—Public-firm patents in races compared to all public-firm patents (2005).** . . . 34
16 **Regression discontinuity setup.** . . . 36
17 **Distribution of title length for leader and follower patents at the time of filing.** . . . 37
18 **Number of patents associated with the customer number (often the attorney) for leader and follower patents at the time of filing.** . . . 37
19 **Histogram showing the incidence of differences in the number of inventors between the leader and follower. The full sample of patented twins.** . . . 39
20 **Line plot showing the percentage of twins by the difference in inventors between the leader and follower. The full sample of patented twins.** . . . 39
21 **Firm patent race winning percentage for the top 100 firms. The simulation curve plots the result of replicating 10,000 times a random draw [win, loss] with 50:50 odds of winning a race for each patent held by the top 100 firms.** . . . 40
22 **Histogram showing the incidence of differences in the number of inventors between the leader and follower. The full sample of patented twins.** . . . 42

23 **Follow-on self-citations.** . . . 45
24 **Patent race leaders do follow-on research that is more similar to their winning patent, than do followers of their losing patents (Right-tail reflects divisionals / continuations).** . . . 46
25 **Representation of the logic used to identify patent transfer type.** . 59
26 **The density of patent transfers by the number of patents included in the transfer.** . . . 63
27 **Linear relationship between examiner score and the number of patents in a transfer.** . . . 63
28 **Percentage of patents transferred. Small transfers include 25 or fewer patents.** . . . 63
29 **Probability that a patent is sold as a function of examiner score, by industry, predicted based on logistic regression estimates.** . . . 71
30 **Probability that a patent is sold as a function of industry-reported product patent effectiveness described by Cohen, Nelson & Walsh (2000), by examiner score, predicted based on logistic regression estimates.** . . . 73
31 **Probability that a patent is sold as a function of market characteristics, by examiner score, predicted based on logistic regression estimates.** . . . 76
32 **Probability that a patent is transferred via a firm exit or a spinoff as a function of market characteristics, by examiner score, predicted based on logistic regression estimates.** . . . 77
B1 **Density of years between patent filing and first post-grant transfer.** 93
B2 **Density of firm size on the patent level.** . . . 93
B3 **Patent transfers by type as a percentage of all transfers, with initial transfers from inventors to firms omitted.** . . . 98
B4 **Density plot of raw similarity values, from Younge and Kuhn (2015).** 99
B5 **Density plot of transformed raw similarity values above 10, from Younge and Kuhn (2015).** . . . 100
B6 **Patent-level density of plot of the maximum similarity value to a previously-filed patent.** . . . 101
B7 **Relative weight of *tf-idf*) terms in a random sample of patent vectors.** 102

# Chapter 1

## Introduction

Patents can play an important role in providing incentives for innovation, and patent strategy is a crucial part of firm strategy for many businesses. However, investigating the effect of patents on firms has long presented significant challenges that stem from fundamental characteristics of innovation and the patent system. An innovation, patent, or technology is by definition different than any other, rendering the comparison of different patents or ideas quite difficult. Moreover, ideas are not randomly assigned to firms or inventors, but rather arise endogenously from a variety of firm, inventor, industry, and technology characteristics. For these reasons, measurement and causal inference are twin problems that hinder research on the effects of patents on firm decisions and outcomes.

This thesis makes several contributions to the literature devoted to overcoming these challenges. First, it describes novel data sets constructed by the author to allow more detailed observation of innovation antecedents, processes, and outcomes. Second, it develops novel measurement tools based on these data sets. Third, it employs novel techniques for drawing causal inferences about the effects of patents on firms. Fourth, it combines these new data sets, measurement tools, and causal instruments to investigate substantive issues of interest to managers and innovation scholars. A central theme of this thesis is that a patent's impact is defined largely by a combination of the characteristics of the underlying idea and the scope of the legal right to exclude afforded by the patent.

Chapter 2, co-authored with Kenneth Younge, describes the construction and validation of the patent-to-patent textual similarity measure. Measuring the technological positioning of a patent relative to other patents is a key antecedent of measuring such diverse concepts as knowledge flows, patent impact, and familiarity. In the past, empirical studies generally adopted one of two methods for positioning patents within a technological space: 1) manual classification of patents by the patent office (e.g., Jaffe 1986, McNamee 2013), and/or 2) linking patents through prior art citations (e.g., Podolny and Stuart 1995, Singh and Agrawal 2011). Although scholars have attempted to improve these approaches, both suffer from intrinsic selection effects in that they rely on human beings (e.g., patent attorneys, inventors, or USPTO employees) to properly assign technological classification or select prior art citations. Both approaches also suffer from the inherent drawbacks associated with being discrete measures necessarily constructed at a relatively fixed level of granularity.

To avoid these problems, we construct a similarity measure by applying a vector space model, one of the most robust methods in the field of Information Retrieval (Manning et al. 2008), to automatically compare the full technical description of every patent to the full technical description of every other patent. Our application of the vector space model builds on the pioneering work by Adam Jaffee (1986, 1989), who constructed a firm-level vector space model using groups of patent classes as dimensions. Taking advantage of advances in computing power, we treat every word that appears in any patent as a separate dimension, yielding more than 700,000 dimensions, and then position more than 5.2 million patents in this space using a term frequency inverse document frequency approach. Finally, we calculate the similarity between every pair of patents using a cosine distance measure.

We show that the text-based similarity measure provides greater accuracy, specificity, and generality than the patent classification system for many research questions. We also show that the patent-to-patent textual similarity measure is much more granular than existing measures of technological relatedness, allowing the scholar to discriminate between patents in the immediate space around a focal patent. Likely for these reasons, the text-based similarity measure proves superior to measures based on the patent classification at predicting phenomena such as patent citations. Finally, we demonstrate how a continuous measure of patent-to-patent similarity can shed light on a wide range of problems in economics and management.

Chapter 3, co-authored with Neil Thompson, discusses patent racing. Patent racing has long been studied in theory but is associated with little empirical analysis. Moreover, although the theoretical work individually produces clear economic predictions regarding the effects of winning a patent race, collectively it produces strongly divergent predictions with results that depend heavily on initial conditions, market power, information availability, and other such factors. The empirical analysis of patent races has long been limited by a host of challenges, including the difficulty of identifying patent races at all, the challenge of identifying winners and losers, the endogeneity inherent in winning or losing a patent race, and the difficulty in measuring the effects of winning or losing.

This chapter introduces a novel approach for identifying patent races as well as a novel instrument for the effect of winning a patent race developed by analyzing patent examination records. We also develop a novel measure of follow-on innovation based on the patent-to-patent textual similarity data. Finally, we demonstrate novel measures of patent impact based on patent citation data.

The analysis shows that many firms are simultaneously "racing" to patent very similar ideas. Unsurprisingly, the winner of a race usually receives a broader patent. However, a firm that exogenously loses a patent race typically also receives a patent, albeit one that is narrower in scope than the winner. Because large firms are engaged in many races and typically win and lose races in equal numbers, the financial effects of winning any particular race are effectively zero because they are usually balanced out by losing a different race. Nevertheless, a patent race does exert a localized effect, providing the winning firm with a more impactful patent and causing the winning firm to conduct more follow-on research.

Chapter 4 discuss the effect of receiving a broader patent on the likelihood that the patent is sold. Patent sales are a critical path by which many firms profit from innovation, and the scope of a patent's legal right to exclude is its defining characteristic. Nevertheless, the empirical analysis of patent sales has been limited by a number of barriers, the first being endogeneity. A patent is an idiosyncratic product of the firm that produces it, tied to technological, economic, and administrative factors. Thus, identifying the causal impact of patent characteristics on an outcome such as patent sales presents a significant challenge. The second barrier is measurement. Patent sales are generally private transactions, with minimal publicly available data regarding the parties, prices, and transferred patents. Moreover, the literature provides few measures for explanatory factors such as patent familiarity and novelty.

I overcome the challenges of measurement by developing a novel data set on patent sales by analyzing patent transfer records. I also use the patent-to-patent textual similarity measure to develop new measures of technological familiarity and technological novelty. Finally,

I describe a new measure of the scope of a patent's legal right to exclude developed in related work by analyzing patent examination records. The related work also extends this new measure to develop a causal instrument for a patent's scope, thus addressing the endogeneity challenges.

I show that an exogenous increase in the legal right to exclude afforded by a patent increases the likelihood that the patent is sold, but that the magnitude of the effect depends on firm, industry, and market characteristics. In particular, an exogenous increase in patent scope has an outsized effect on patent sales for small firms and for firms in industries where patents are reported as being effective for exclusion. In keeping with the literature on market design, an exogenous increase in patent scope has a large effect in thick and uncongested markets and virtually no effect in thin or congested markets.

Thus, Chapter 2 describes and validates data that is used to construct measures of patent impact, novelty, familiarity and rights congestion that supports the analysis in Chapters 3 and 4. Chapter 3 shows that a significant business-related *cause* of receiving broader or narrower patent scope is a firm's standing relative to other firms in a patent race. Chapter 4 demonstrates some of the *effects* of broader patent scope on firms outcomes and decisions.

# Chapter 2

### Patent-to-patent Similarity: A Vector Space Model

## Introduction

Concepts of technological space, distance, and relatedness are central to the study of invention and innovation. As Arrow points out, "the determination of optimal resource allocation for invention will depend on the technological characteristics of the invention process and the nature of the market for knowledge" (Arrow 1962). Researchers rely on technological space, explicitly or implicitly, to model a wide-range of phenomena, including technological diffusion (Nelson and Phelps 1966), technological opportunity (Rosenberg 1974), endogenous growth (Romer 1990), geographic spillovers (Jaffe, Trajtenberg et al. 1993), research productivity (Henderson and Cockburn 1996), knowledge transfer (Mowery, Oxley et al. 1996), corporate diversification (Silverman 1999), employee mobility (Almeida and Kogut 1999), boundary spanning (Rosenkopf and Nerkar 2001), knowledge recombination (Fleming 2001), knowledge complementarity (Nesta and Saviotti 2005), alliance networks (Gilsing, Nooteboom et al. 2008), competitive rivalry (Bloom, Schankerman et al. 2013), and many others. As technology is difficult to observe on a broad scale, much scholarship relies on the patenting system as a proxy (Griliches 1979).

Empirical studies generally adopt one of two methods for positioning patents within a technological space: 1) manual classification of patents by the patent office (e.g., Jaffe 1986, McNamee 2013), and/or 2) linking patents through prior art citations (e.g., Podolny and Stuart 1995, Singh and Agrawal 2011). For example, research on knowledge spillovers combines both methods by examining patent citations, while controlling for technology classification (Jaffe, Trajtenberg et al. 1993). Researchers have attempted to improve these methods by moving to lower levels of granularity (Thompson and Fox-Kean 2005), incorporating multiple classifications (Benner and Waldfogel 2008), adjusting for relative distances within the classification taxonomy (McNamee 2013), and/or switching to alternative schemes such as the International Patent Classification system. Although such efforts have merit, concerns persist "about the accuracy of the more detailed U.S. patent subclasses" (Belenzon and Schankerman 2013) and one wonders if the field is beating a dead horse (or at least one that does not run in the way hoped for by researchers). For their part, patent attorneys indicate that they now largely ignore the classification system in favor of text-based searching.

Scholars working with patent data often need to identify, compare, and/or match patents from similar domains of technology (Singh and Agrawal 2011, for example, use matching methods to compare firms and inventors with similar patent portfolios). Several problems arise from using the patent classification system for such purposes. First, patent classes may not be independent from the phenomena of interest. Patent classes are created, grouped, merged, split, and/or reassigned in ways that can correlate to changes in the technological space, and such effects can introduce a sample selection bias into the analysis. Second, patent classes may be too broad for the types of counterfactual comparisons attempted by researchers. For example, Schmookler observed patents for both a toothpaste tube and a manure spreader in the same "dispensing of solids" patent class/subclass (Schmookler 1966: P. 20, Jaffe 1986). Similarly, patents for a compact catheter assembly (U.S. Patent

No. 8,556,884) and a device for practicing golf (U.S. Patent No. 5,853,334) are located in class/subclass 242/370 because they both involve "winding, tensioning, or guiding." Third, patent classes may at times be too precise, leading to a curse of dimensionality that lowers recall and the ability to find relevant comparisons (Benner and Waldfogel 2008). Narrow subclasses can also lead to measurement error as the USPTO attempts to pick the "best" subclass out of 150,000 options (Henderson, Jaffe et al. 2005). Overall, the econometric problem is that categorical classification can select systematic differences into a sample, which then are difficult to identify through sensitivity and robustness tests. As such, this chapter acts on the argument "the best way forward is to devise identification systems that avoid entirely technology classification systems" (Thompson and Fox-Kean 2005b: Page 466).

This chapter demonstrates that the United States Patent and Trademark Office (USPTO) and World Intellectual Property Organization (WIPO) patent classification systems are too coarse for many of the comparisons sought after in academic research. Although this critique of the classification system is not new, the analysis in this chapter differs from previous work by leveraging information retrieval theory to develop a machine-automated measure of patent-to-patent similarity (one that is independent of institutionalized systems of classification) as a potential solution to the problem. Moreover, the results show that a text analysis of the technical description (i.e., "specification") can discriminate between patents in the immediate space around a focal patent. In contrast, patent classes and subclasses include a considerable range of diversity, such that patent classes may be more appropriate for identifying broad, functional categories, as opposed to technological similarities. The proposed measure is also an initial step toward exploiting Big Data methods in patent research methods that hold great promise for analyzing phenomena of high dimensionality (Fan, Han et al. 2014).

This chapter make several contributions related to technological similarity. First, it introduces and validates a continuous measure of technological similarity; prior research has largely relied on discrete divisions between categories for such measures. Second, the continuous measure is calculated at the dyadic, patent-to-patent level across the entire population of patents, whereas prior research has relied on small subsets of data, or "profile" measures aggregated up to the level of the firm. Third, the analysis demonstrates that a text-based similarity measure provides greater accuracy, specificity, and generality than the patent classification system for many research questions. Fourth, the analysis demonstrates how a continuous measure of patent-to-patent similarity can shed light on a wide range of problems in economics and management.

## The Vector Space Model

Vector space models are one of the most robust methods in the field of Information Retrieval (Manning, Raghavan et al. 2008). Originally developed for the SMART retrieval system (Salton, Wong et al. 1975), vector space models now motivate nearly every modern search engine (Turney and Pantel 2010). This project maps the entire USPTO patent space into a single vector space model at the patent-to-patent level.

Jaffe (1986, Jaffe 1989) pioneered the use of vector space models at the level of the firm to position firms based on their patent portfolios. The 'Jaffe Method' constructs a profile for each firm, as vector F = (F1...FK) across K patent class dimensions, where Fk equals the

fraction of patents held by a firm in the k th patent class (e.g., Ahuja 2000, Song, Almeida et al. 2003). Given a vector for each firm, the Euclidean distance (e.g., Rosenkopf and Almeida 2003), angular separation (e.g., Agarwal, Ganco et al. 2009), correlation (e.g., Benner and Waldfogel 2008), min-complement (e.g., Bar and Leiponen 2012), or Mahalanobis distance (e.g., Bloom, Schankerman et al. 2013) can then be calculated between any two firms.

While the above methods build on a "vector space model," they have important limitations. First, prior research relies on a discrete categorization within the patent classification system to define the dimensionality of the model (e.g., Chen and Chiu 2011). This chapter also examines problems that arise from using the classification system and discrete comparisons to measure technological distance. Second, prior research aggregates data to the firm. Aggregation to the firm discards useful information at the patent level and conflates firms of different sizes, diversification, focus, and other characteristics. Moreover, the combination of categorization and firm-level aggregation leads to difficult tradeoffs in determining the optimal distance metric to use for a given research question (see Table X in Bloom, Schankerman, Van Reenen, 2013: page 1388).

Section 4 describes the calculation of the similarity of every possible patent-to-patent similarity comparison, for over 14 trillion calculations (an admittedly brute-force approach). The dimensionality of the model is constructed from the vocabulary used in patent documents, rather than the assignment of patent classifications. Doing so allows us to capture the similarity between technological fields where patent classifications do not overlap, a key concern raised by Bloom et al. (2013). Moreover, lower-level data allow researchers to explicitly construct, and then aggregate, measures up to the firm level to address the particular research questions of their study. While there have been earlier attempts in this direction (e.g., Nanda, Younge et al. 2015), studies to date have focused on small portions of the overall space and have lacked comprehensive validation.

In the field of information retrieval, vector space models represent a high-dimensional space based on the full text of a given sample (or "corpus") of documents. The corpus is the complete set of utility patents granted by the USPTO from 1976 through 2014. While most research samples from a population, the analysis described in this chapter calculates the pairwise similarity between every member of the corpus, for a full population analysis. HTML was scraped for each granted patent from the USPTO website and converted to plain text. Then, the section for the technical description was extracted. Scraping from patent number 3,552,244 to 8,952,843 (a set of 5,400,599 potential patents), the analysis includes 5,298,356 patents (98%). Spot checking suggests that missing observations were withdrawn patents, missing from the USPTO system, or malformed records.

Using a 'bag of words' methodology (Turney and Pantel 2010), the VSM dimensions are definedbased on the vocabulary of terms used by patents within the sample. The technical description for each patent was tokenized, punctuation was removed, variations were stemmed with an English Porter2 ("snowball") algorithm, and terms truncated at 25 characters. Terms with numbers, genetic sequences, a high proportion of repeating characters, natural-language stop words, legal stop words identified through interviews with patent attorneys, and terms with less than 3 characters were dropped. Finally, we drop terms with extremely high document-frequencies and terms appearing in fewer than 10 patents (misspellings or unnatural language). Overall, iterative filtering helped to optimize the balance between Type-II errors (i.e., missing overlap), Type-I errors (i.e., erroneous overlap), and

efficiency (Ming, Wang et al. 2010).

After building the vocabulary, a weighted vector was generated for each patent into the "space" based on the term-frequency of each term for a patent, scaled by the inverse document-frequency of each term across the corpus. Called term-frequency-inverse-document-frequency (Aizawa 2003), the *tf-idf* method allocates more importance to terms used frequently within a patent, while attenuating the importance of terms appearing widely across the corpus. In other words, the weight of a term for a given patent represents the product of the importance of the term for the patent, and the amount of information embedded in the term as determined by the corpus. *Tf-idf* weighting is one of the most widely used methods in text analysis; theoretical work suggests it should be efficient and consistent for most retrieval tasks (Aizawa 2003, Wu, Luk et al. 2008), and empirical work suggests it works well in practice (Elkan 2005).

**Table 1: Summary statistics.**

**a.** Source data.

| | | |
|---|---|---|
| Patents | Total count | 5,298,356 |
| Patents | Date range | 1976 through 2014 |
| Patents | Numerical range (USPTO patent numbers) | 3552244 through 8952843 |
| Vocabulary | Count of terms (unique) | 6,333,830 |
| Vocabulary | Count of terms used in VSM (dimensions) | 743,253 |

**b.** Sample construction.

| | (1) | (2) | (3) |
|---|---|---|---|
| | Full Population* | Retained Calculations | Final Sample |
| Observations | 14.0 trillion | 370.8 billion | 14.0 trillion |
| Percent of full pop. | 100% | 2.73% | 100% |
| Data storage | 296.3 TB | 8.1 TB | 8.1 TB |
| Similarity – Mean | 0.0177 | 0.1748 | 0.0051 |
| Similarity – S.D. | 0.0335 | 0.0844 | 0.0331 |
| Similarity – Skew | 5.87 | 2.28 | 7.80 |

To measure similarity, the cosine of the angular separation was calculated for every two patents in the population. Figure 1 illustrates a simple example in three dimensions: vector A and vector B point out from the origin, $\theta$ is the angular separation between A and B, and cosine($\theta$) measures the "similarity" of the two vectors. The cosine of $\theta$ varies from zero (completely orthogonal) to one (completely overlapping). As reported in Table 1, most

calculations for the full population were in fact close to zero (Mean=0.0177), and 97.27% of calculations had a raw cosine similarity score of less than 0.10. Manual review of similarity scores less than 0.10 indicated that variation below that threshold was essentially noise and that such measurements could be discarded from the dataset (i.e., not stored) and imputed later. Dropping data below a score of 0.10 reduced the storage and query requirement from 296.3 TB to 8.1 TB.

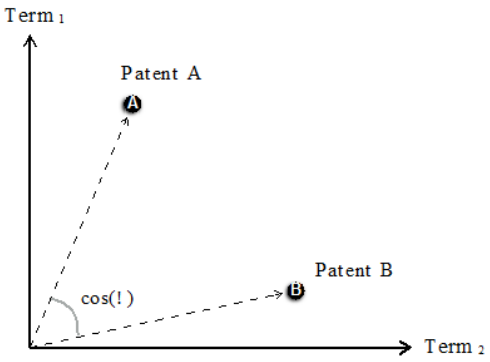
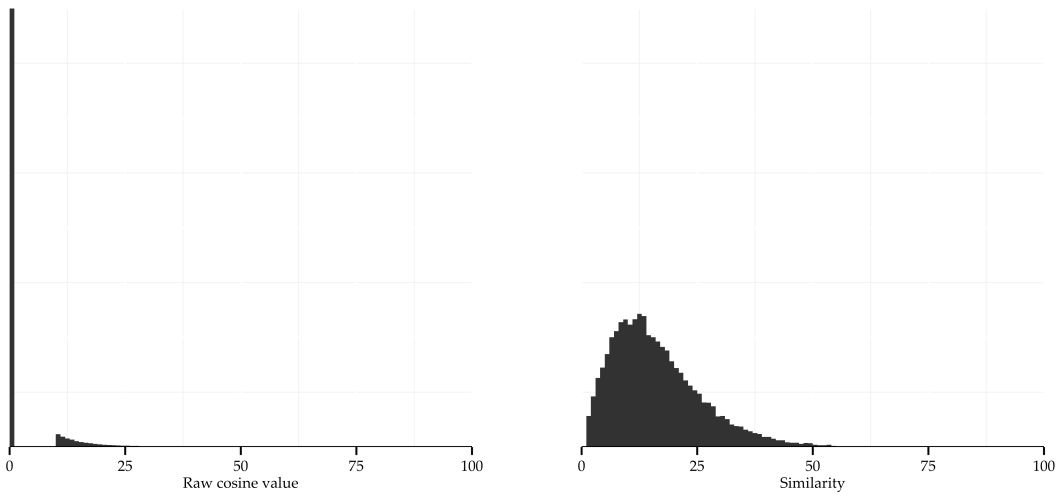



**Figure 1: Angular separation in a vector space model.**




**(a) Full population (n=14,036,290,800,546)**

**(b) Final Sample (n=370,842,517,085)**

**Figure 2: Distribution of similarity by sample.**

Finally, although it is tempting to think of similarity as a percentage, manual inspection suggests that cosine similarity becomes informative around 0.20, and that two patents are quite similar by 0.50. Results were then transformed by stretching them from the captured

0.10 - 1.0 interval back to the 0.0 - 1.0 interval, and then radially stretching results away from the ends of the interval while maintaining rank order. Summary statistics for the original Full Population, and transformed Final Sample are presented in Table 1 and graphed in Figure 2.

## Validation

This section validates four important aspects of the patent-to-patent measure of similarity: internal validity, content validity, external validity, and predictive validity.

**Internal Validity.** Internal validity refers to the degree to which a measure minimizes systematic error (Nunnally and Berstein 1994, Short, Broberg et al. 2010). Systematic error can result from malformed theory (where factors within a construct do not aggregate or interact as expected) or from errors made during computation. This section first reviews the steps taken at both the theoretical level and computational level to avoid such problems.

When defining the vocabulary space, there is a tradeoff between identifying attributes that are rare and therefore informative, and giving too much power to erroneous terms. If the vocabulary space is too small, it will overlook important information (akin to a Type-II error); if the vocabulary space is too large, it will draw false connections (akin to a Type-I error). For example, a rare form of spelling mistake can, once weighted by tf-idf, increase the similarity between patents that share the spelling mistake; this is especially true when the spelling mistake is made on a common term that otherwise has little weight. Alternatively, failure to identify rare but highly distinguishing terms (e.g., "CRISPR"), will fail to identify a connection that could be clear to a human reader. As the optimal solution to this dilemma is difficult to determine a priori, the vocabulary was parsed, cleaned, and constructed through many iterative passes. After each pass (but before calculating the VSM), the vocabulary was inspected, tested, and recalibrated. Future work may benefit from a posteriori analysis of competing vocabularies, but budget limitations prevented us from making multiple passes at the entire similarity calculation itself.

A truly Big Data calculation (273 TB; $n = 14,036,290,800,546$), the procedure was distributed over thousands of computers running on the Google Cloud platform for many weeks. Errors can arise for many reasons, including machine failures, network failures, and programming errors. When that happens, small ranges of data go missing. If processing is ordered by a variable of interest (such as time, as it often is for computational efficiency), then missing data will no longer be 'missing at random,' and missing data can bias subsequent results. One cannot visually scan Big Data for errors, so additional steps were taken to verify the results during and after computation. Computational validity was also checked throughout the pipeline by verifying that counts and distributions of output data were consistent with the values expected from the input data.

**Content Validity.** The text analysis approach used for the vector space model derives directly from the content of patents themselves, and as such one would expect the content validity of the VSM to be intrinsically high. Nevertheless, several face validity checks were performed to ensure that the computed similarity values reflect real world phenomena. First, the VSM similarity within a sample rises as one moves from a random sample to samples from companies with a narrower range of a given technology. Second, the average VSM similarity of patent "families" is particularly high. Given that patent families share some or

all of the text of the technical description amongst family members (through "patent priority claims"), one would expect this to be the case.

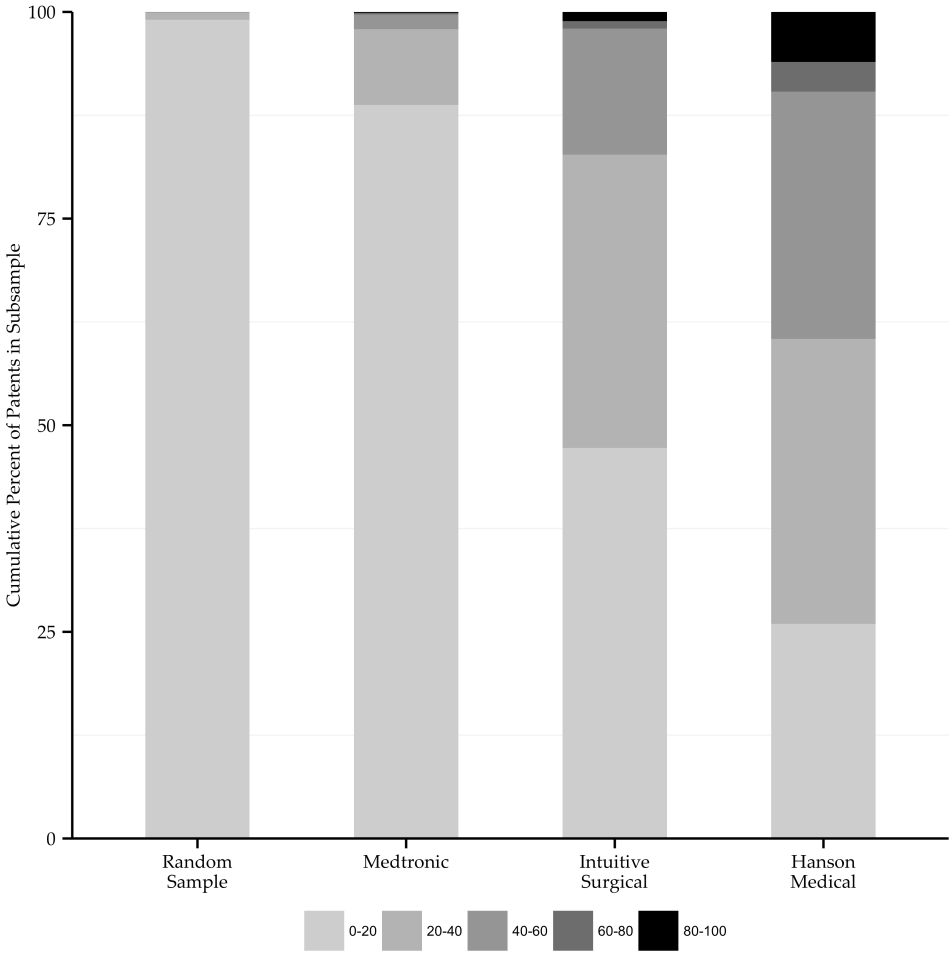



**Figure 3: Patent similarity within firms.**

The first test of face validity was whether VSM similarity rises as one examines the patent portfolios of increasingly focused firms. Figure 3 plots the pairwise VSM similarity between every two patents held within several firms in the medical instruments industry. We start with a random sample, and then move through firms known to have an increasing focus on a narrower and narrower slice of the medical instruments space. Starting with a random sample, the great majority of pairwise comparisons ($> 97\%$) share essentially no similarity (i.e., they have a similarity score $< 0.20$). Next, moving to Medtronic - a large medical instruments company with R&D in many different areas of the medical instruments space - the proportion of patents held by the firm with a pairwise similarity score of 0.20 to 0.40 rises, and a small share appears at even higher levels of similarity. Next, moving to Intuitive Surgical — a midsized company focused on robotically assisted surgery — the share of unrelated patents continues to fall, the share of somewhat related patents continues to expand, and the share of highly related patents ($> 0.40$) continues to expand. Finally,

moving to Hansen Medical — a small and tightly focused company that patents exclusively in intravenous, catheter-based surgical systems — the share of unrelated patents continues to fall, the share of related patents continues to grow, and the share of highly-related patents (i.e., pairwise patents with a similarity score $> 0.40$) continues to grow to be the largest of the samples. The rising distribution of similarity as one moves from less-to-more focused R&D activities face validation for the content validity of the VSM measure.

A second test of content validity is whether VSM similarity scores higher when one knows that the actual content of two patents should be more similar. This test examines "patent families," a phenomenon driven by patent priority. An applicant may file an application with a priority claim to expand upon claims filed in an earlier-filed application, while removing some prior art from consideration that would otherwise block the patent application. Importantly, a priority claim is valid only insofar as the later-filed application claims an invention that was already disclosed by the earlier filed application. Thus, two applications linked by a priority claim should be highly similar.

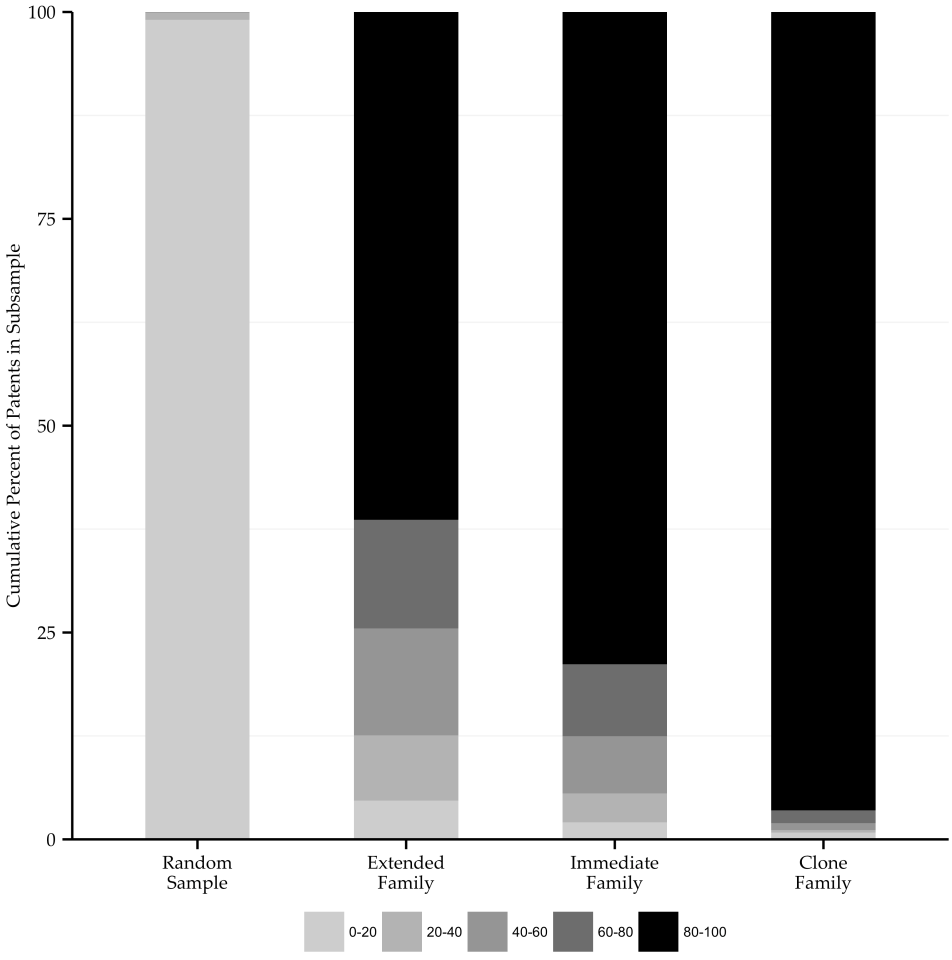



**Figure 4: Patent similarity within patent families.**

For this analysis, a patent family all patents that are linked by a U.S. patent priority

claim. Patents with the most restrictive priority claims, referred to as continuation or divisional patents, should be virtually identical to their earlier-filed parent applications. These are termed "clones." A somewhat less restrictive grouping includes clones, but also pairs of patents that share a priority claim to a common provisional patent application. These are termed "immediate families." The loosest type of priority relationships are those established by a "continuation-in-part," in which the application declares that the later-filed application substantively expands upon the specification of the earlier-filed application. Groups of patents joined by such linkages are termed "extended families."

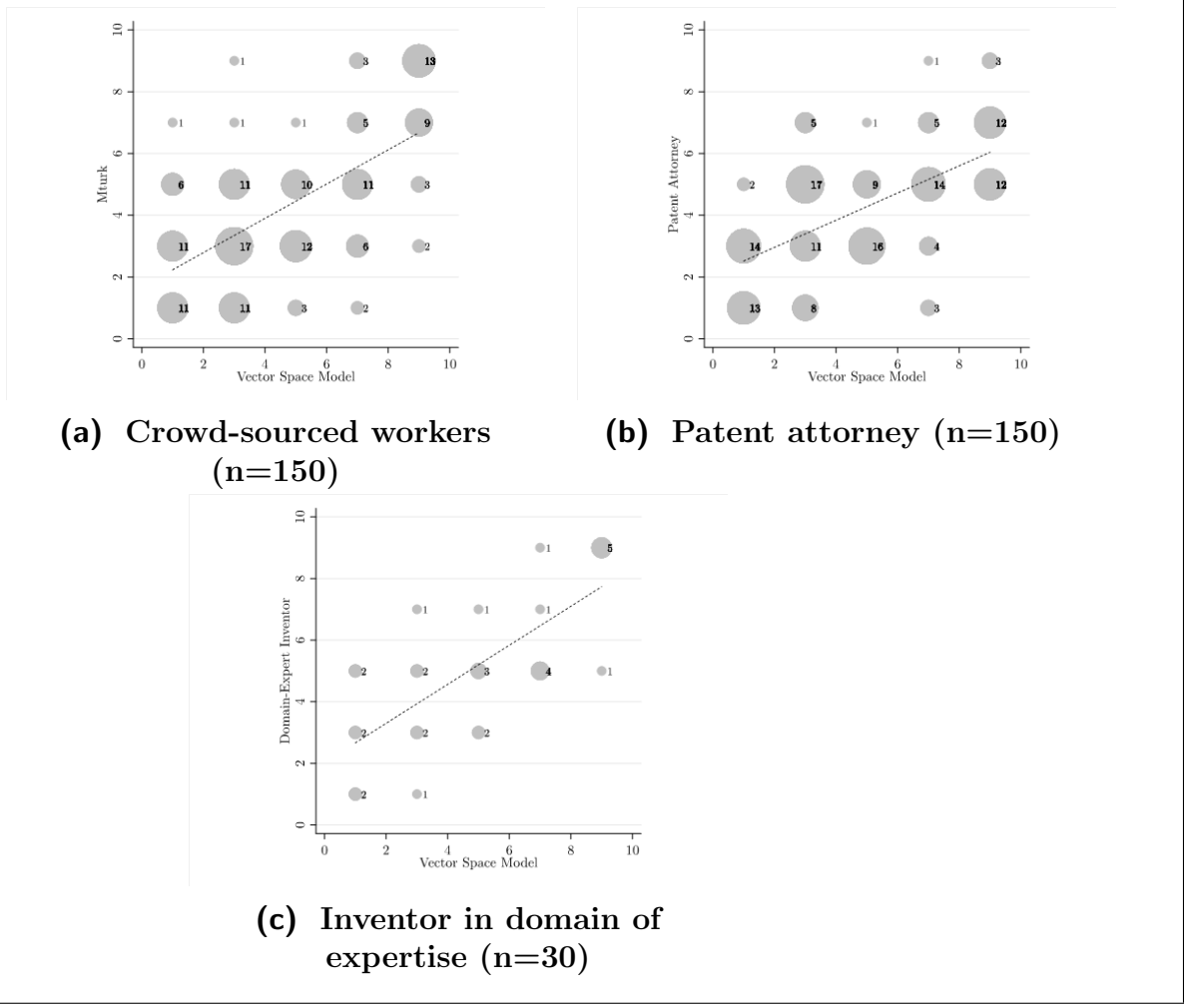
The VSM also correctly identifies family linkages. Figure 4 includes the breakdown of similarity values for random samples of patent dyads that share clone, immediate, and extended family relationships. Virtually all patent pairs linked by continuation or divisional are highly similar, with over 96% having a similarity value of at least 0.99. As expected, the correlation between family relationship and similarity decreases slightly but remains quite high as the definition of families is relaxed to include immediate and extended family members.

Because the similarity measure matches independent phenomena both inside and outside the patent system, the similarity measure seems to exhibits substantial content validity.

**External Validity.** To test the external validity of the VSM against human perceptions, a set of crowd-sourced raters, a U.S. patent attorney, and an expert inventor were recruited. Each was asked to rate a random sample of pairwise-comparisons from the VSM results. Because a simple random sample of pairs from the dataset would not select enough cases of high-similarity to assess the high end of the scale, a random sample stratified was selected at similarity levels of 0.10, 0.30, 0.50, 0.70, and 0.90, with an approximately equal number of comparisons at each level.

First, Figure 5 tests the correlation of the vector space model to individuals who are not experts in the patenting system. Seven workers were recruited from the Amazon Mechanical Turk crowdsourcing system and asked them to rate each of the 150 pairwise-comparisons on a scale of: (9) extremely similar, (7) very similar, (5) somewhat similar, (3) not very similar, and (1) extremely dissimilar. Prior research demonstrates that a measure of central tendency by multiple, independent raters, can generate a surprisingly accurate assessment (Fuxman, Tsaparas et al. 2008). The instructions therefore advertised that the payment would double for ratings that agreed with the consensus rating of other raters assigned to the same comparison. We take the mode rating from seven raters as the consensus rating. Panel A of Figure 5 plots the results, with the number of consensus ratings at a given level of VSM similarity indicated on the graph, the size of each circle plotted proportional to the number of ratings, and the OLS best-fit line is overlaid as a dashed line. The Pearson correlation between the consensus rating and the vector space model is moderate [$\rho = 0.63$, $\chi^2 = 90.24$], the inter-rater reliability is high [Cronbach $\alpha = 0.77$], and the slope of the OLS best-fit line is highly significant [$p < 0.0001$].

Second, to test the external validity of the VSM measure against one skilled in the art of prosecuting patents, an independent US patent attorney was hired to manually rate the same sample of patent comparisons presented to the crowd-sourced workers, using the same scale. Panel B of Figure 5 plots the results. The Pearson correlation between the attorney rating and the vector space model was again moderate ($\rho = 0.60, \chi^2 = 87.96$), the inter-rater reliability to be high (Cronbach $\alpha = 0.75$), and the slope of the OLS best-fit line to be highly significant ($p < 0.0001$). A remarkable consistency exists between results from the




**(a) Crowd-sourced workers (n=150)**

**(b) Patent attorney (n=150)**

**(c) Inventor in domain of expertise (n=30)**

**Figure 5: Probability that a patent is sold as a function of market characteristics, by examiner score, predicted based on logistic regression estimates.**

crowd-sourced and patent attorney validation test. In a three-way test of inter-rater reliability between the expert, crowd-sourced, and VSM rating, the ratings exhibit an overall inter-rater reliability of Cronbach $\alpha = 0.84$, a very high level even for psychometric studies (Cronbach 1971, Cortina 1993, Nunnally and Berstein 1994).

Finally, to test external validity in a particular domain, a an inventor familiar with the technology was recruited to rate a random sample of pair-wise comparisons of patents from Intuitive Surgical Inc. (as described earlier under content validity). The sample was again stratified at similarity levels of 0.10, 0.30, 0.50, 0.70, and 0.90. Time constraints with the (volunteer) inventor prevented us from testing an equivalently sized sample as with the (paid) patent attorney and crowd workers, so the sample size is reduced from 150 to 30. Panel C of Figure 5 plots the results. The Pearson correlation between the inventor and the vector space model increases substantially ($\rho = 0.73$, $\chi^2 = 29.17$), the inter-rater reliability to also

increases (Cronbach $\alpha = 0.85$), and the slope of the OLS best-fit line steepens somewhat ($p < 0.0001$).

The consistency of external validity tests across different types of external raters lead us to conclude that the vector space model generally conforms to what a human rater would expect from the measure. In the post-rating discussions with the attorney and inventor, however, the VSM measure of technological similarity does not necessarily demark the areas of intellectual property that are protected by a given patent, or how one patent might block another. Instead, the measure identifies technological elements in common between two patents.

**Predictive Validity.** Predictive validity is the extent to which a test measure can predict a criterion measure. In patent data, the clearest criterion measure is whether a pair of patents is also classified as similar by patent applicants and/or the USPTO. Patent applicants and/or the USPTO may declare two patents to be similar in one of two ways. First, a previously-issued patent may be cited during the prosecution of a later-issued patent. Second, two patents may be similarly classified by the patent office, either explicitly through the USPTO's patent classification system or implicitly through assignment into the same administrative patent examining division.

We begin by using patent similarity to predict patent citations. McNamee (2013) provides the most powerful test to date for predicting patent citations: he constructs a measure of the degree to which a pair of patents shares the same subclass(es) under the USPTO's patent classification system. To test the measure, McNamee constructed a sample of all USPTO patents from class 704 (either primary class or additional class). He split the dataset into two subsets by grant year: the first subset (1997-2001), was taken as the potential issuer of a citation; the second subset (2002-2006), was taken as the potential receiver of a citation. We duplicate the sample in McNamee and compare the variable VSM Similarity to the subclass overlap variable reported in McNamee.

Table 2 reports predictions of patent citation for the VSM measure of patent similarity. All models were estimated with relogit, a rare events logistic regression package in Stata (King and Zeng 2001). Although citation is a very rare event (less than 0.001 of patent-to-patent pairs), the sample size for the rare event was greater than 16,000, which should be sufficient to avoid small-sample bias in logistic regression. Columns 1 and 2 predict whether a given dyad also exists as a citation (no: 0; yes: 1). Column 1 copies results from (McNamee, 2013) based on his measure of extended patent subclass overlap. Column 2 replicates the sample reported by McNamee, and performs an equivalent analysis using the VSM measure of textual similarity. The aforementioned sample construction resulted in approximately 15 million potential citation dyads for analysis. In Column 1, McNamee finds extended subclass overlap to be a highly significant predictor of citation, with subclass similarity predicting 12% of variation. Column 2 shows that the VSM similarity measure is also a highly significant predictor of citation, explaining 14% of variation.

Table 27 indicates that the vector space model outperforms subclass similarity, even when looking within a sample of dyad pairs pre-selected to share the same main class. Selecting a sample where patents share the same class, and then using a class-based measure to predict citations, is also a form of selecting on the dependent variable and likely overstates the power of the taxonomy. The analysis therefore contradicts McNamee's conclusion that "the type of analyses done in dataset #1 are not possible without taxonomically appropriate methods like

those presented in this paper" (McNamee 2013: p. 871). Additionally, an important limitation of relying on the patent classification system is that one cannot analyse the strengths or weaknesses of the patent classification system itself. An advantage of the VSM similarity measure is that it is not dependent upon any explicit classification scheme; it works with a randomly selected sample just as well as with a sample selected based on patent class. Therefore, the next section applies a regression framework where using patent similarity to predict patent classification.

**Table 2: Predictions of patent citation.**

| | (1) | (2) |
|---|---|---|
| | McNamee | VSM |
| Extended Subclass Overlap | 0.62*** | |
| | (0.00) | |
| Similarity | | 7.7183*** |
| | | (0.035) |
| Constant | -10.08*** | -8.2060*** |
| | (0.02) | (0.013) |
| Total patent-to-patent pairs | 15,126,093 | 15,741,565 |
| Cited patent-to-patent pairs | 16,285 | 16,221 |
| Pseudo $R^2$ | 0.12 | 0.14 |

Table 3 analyzes whether similarity predicts an overlap in patent classification. The sample is drawn from a random sample of 16,000 patents. After constructing a unique list of all pairwise combinations between the 16,000 patents (i.e., the bottom triangle of the cross-joined matrix), the sample resulted in 1,675,563,971 observations. The transformed Vector Space Similarity variable was merged in from the vector space model, and indicator variables were merged in from USPTO metadata to indicate whether each pairwise comparison was assigned to the same Art Unit for examination (DV in Column 1), shared a primary USPTO main class (DV in Column 2), shared a primary USPTO main class and subclass (DV in Column 3), or shared an Art Unit, primary USPTO main class and subclass (DV in Column 4). The pseudo R2 for each model was calculated using the relogit function in Stata. As is shown in Table 3, the VSM measure is a strong predictor of shared main class (Column 1) and of main class and subclass (Column 2). The transformed VSM variable is highly significant at the p=.001 level, and explains approximately 19% of the variation in whether or not a patent dyad shares the same primary USPTO main class and subclass. The predictive validity improves even more when predicting dyads that not only share the same primary main class and subclass, but are also grouped by the USPTO into the same Art Unit. In this case, the VSM measure explains 22% of the variation.

Based on the results reported above, a vector space model based on the technical description of patent applications appears to be a reliable and consistent measure of similarity. First, the VSM measure is a highly significant predictor of pre-existing measures of patent similarity such as shared main class or subclass. Second, the VSM measure explains a greater

proportion of the variation than pre-existing measures of patent similarity. Third, the VSM measure is not reliant on human-based classification schemes such as patent class and thus is not subject to selection effects. Fourth, by being independent of human-based classification schemes, the VSM measure can be compared with existing classification schemes without selecting on the dependent variable. Further research is needed to assess the full merits of the measure for applied work, but initial tests suggest that it may be a powerful and broadly applicable tool for understanding the similarity and relatedness of patents across a range of contexts.

**Table 3: Predictions of patent classification.**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable (No=0 / Yes=1) | Shares Art Unit | Shares Main-class | Shares Sub-class | Shares All |
| Vector Space Similarity | 10.3161*** | 11.6959*** | 12.3782*** | 12.9403*** |
| | (0.010) | (0.008) | (0.025) | (0.041) |
| Constant | -5.9504*** | -4.9455*** | -9.2609*** | -10.6254*** |
| | (0.001) | (0.001) | (0.007) | (0.014) |
| Total patent-to-patent pairs | 163,004,589 | 163,004,589 | 163,004,589 | 163,004,589 |
| Overlap patent-to-patent pairs | 542,882 | 1,527,228 | 26,967 | 7,922 |
| Pseudo-$R^2$ | 0.08 | 0.10 | 0.16 | 0.20 |

## Classification

This section considers how automated measurements of similarity by a vector space model compare to manual classification of patents by the patent office. Given that the dominant method for scholars to position patents in a technological space relies on the manual classification of patents by the patent office (Benner and Waldfogel 2008), it is important for research to assess and understand the relative strengths and weaknesses of the classification system. As reported below, the VSM is better suited for identifying near-neighbors, while the patent classification system is better suited for identifying broad, categorical groupings.

Recently, scholars have highlighted the fact that most patents are actually cross-classified into many different classes and subclasses (Benner and Waldfogel 2008, Bar and Leiponen 2012, McNamee 2013), not just the primary one referenced in most studies. The NBER patent data file (Hall, Jaffe et al. 2001), for example, drops multiple classifications altogether and retains only the primary classification for analysis. Given that the NBER patent file is the most frequently used patent dataset, it is perhaps unsurprising that most studies do not address multiple classification assignments. To explore the issue, we rely on a parsing of the XML metadata files from the USPTO from 2005 through 2014. We identified 491 distinct USPTO main classes and 82,520 distinct subclasses at the six-digit level. (Henderson, Jaffe et al. 2005). For patents granted after 2005, 20.1% of patents were assigned to just one classification (the "primary" main class / sub class combination), 23.4% were assigned two classifications, and 56.5% were assigned to three or more classifications.

While early research focused on the primary main class (e.g., Jaffe, Trajtenberg et al. 1993), later research argued that main class alone may be too coarse (e.g., Thompson and Fox-Kean 2005). While some research has moved to use subclasses (e.g., Belenzon and Schankerman 2013), others have argued that subclasses may increase measurement error (Henderson, Jaffe et al. 2005). Yet other work suggests that measures of overlap for one subclass suffers small-sample bias, such that scholars should use the secondary (multiple) subclass classifications to cast a wider net and reduce sampling variation (Benner and Waldfogel 2008). And Bar and Leiponen (2012) point out that multiple classifications can be irrelevant when making portfolio-to-portfolio comparisons.

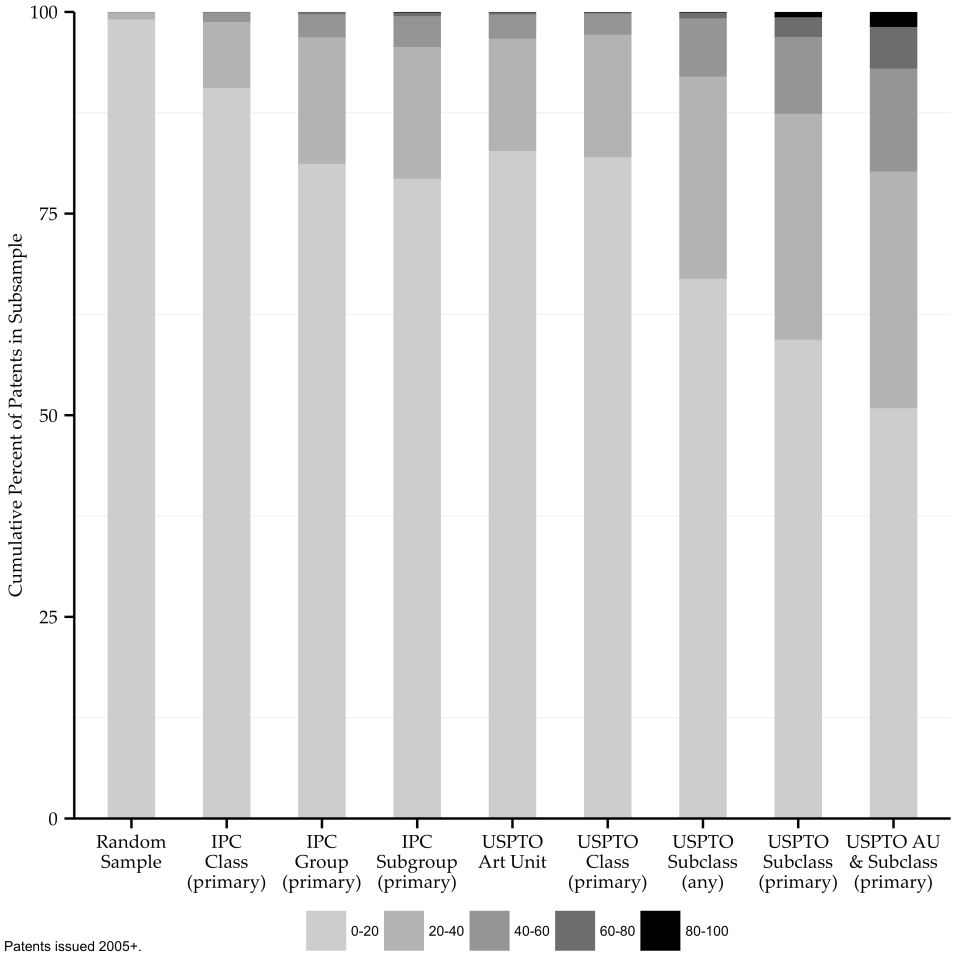
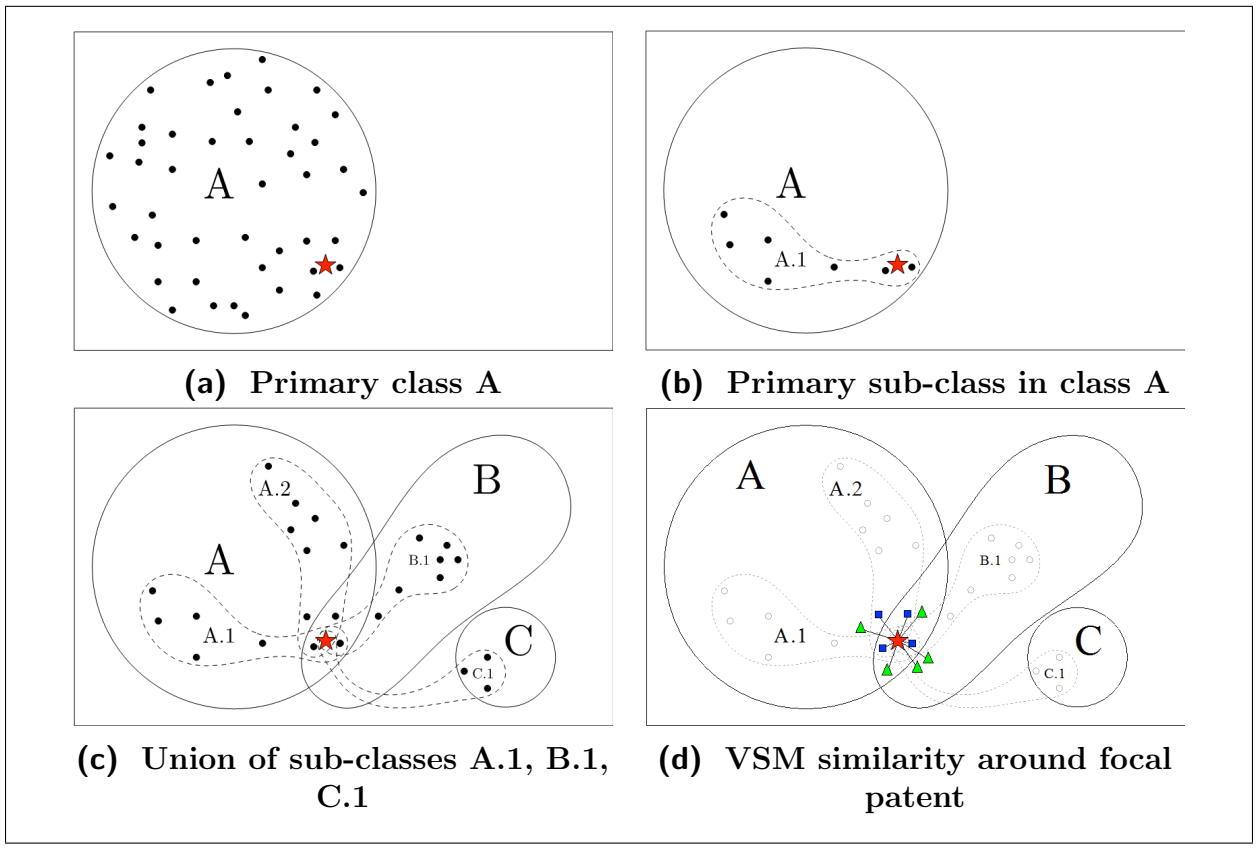



**Figure 6: Patent similarity within patent classifications.**

To better understand the relative merits of using subclasses, we first examine the distribution of similarity for patents that share such categorization. Figure 6 plots the cumulative percent of pairwise comparisons (where both patents fall within the same classification) for given ranges of similarity. A random subsample was constructed for each column where each patent pair in the sample shares the grouping indicated in the figure. A random sample from across the entire dataset is included for comparison. For the sample of dyads that share the

same USPTO main class and subclass, the results show that 30% of dyads fall below a similarity value threshold of 0.20, and only 10% of dyads have similarity values of greater than 0.40. Viewed another way, a plot of the cumulative distribution of patent-to-patent pairs within a shared primary USPTO subclass indicates that over one-third of patents within a primary USPTO subclass are too dissimilar to be accurately rated by the vector space model, resulting in similarity values of approximately zero.

Earlier in this section, results were reported from a logistic model using the VSM measure of patent text similarity to predict patent classification at the main class plus subclass level; that model explained 19% of the variation in patent class assignment. The unexplained variation, however, is not necessarily due to measurement error or inconsistency in the VSM predictor variable. Instead, most dyads within a given patent classification simply have a low level of similarity. Due to the *tf-idf* weighting of vectors, the VSM measure is better suited for identifying differences between near-neighbours than it is suited for discriminating between distant relations, in part because relatively few key terms determine the overall variation in the scale.




**(a) Primary class A** **(b) Primary sub-class in class A**

**(c) Union of sub-classes A.1, B.1, C.1** **(d) VSM similarity around focal patent**

**Figure 7: Overlapping patent classifications.**

In all of the empirical results, the measure of similarity exhibits considerable variation within subclasses. The question therefore arises as to whether the patent classification system can be exploited in some way to accurately center around a patent of interest. To illustrate the point, Figure 7 presents four Venn diagrams to illustrate various problems that arise from

overlapping patent classifications. A focal patent of interest is represented by a red star, and other patents are distributed as black dots in spatial relation to their VSM similarity. Panel A illustrates the full set of patents associated with a hypothetical primary patent class A, and panel B illustrates the set of patents associated with sub-class A.1. The problem here is that a patent near the boundary of the class can actually have a lower similarity to other members of its subclass, as compared to some members of its main class. Furthermore, Panel C illustrates that taking the set of patents formed through the union of sub-classes A.1, B.1, and C.1, as suggested by Benner and Waldfogel (2008), can increase the problem. By comparison, Panel D illustrates that a multidimensional set of patents, as identified by patent similarity, can potentially improve the selection by identifying both new patents (plotted as green triangles), while including some previously identified patents (plotted as blue squares), and while excluding distant/irrelevant patents. Although the exact mix and number of new patents, retained patents, and excluded patents will depend on the coarseness and accuracy of the patent classification system in a particular domain, the illustrations of Figure 7 demonstrate the potential for systematic error in the classification scheme.

To the extent that a vector space model cuts across arbitrary selection boundaries, the VSM should generate a more precise—and more reliable—measure of technological space. Nevertheless, the validity of patent classification depends on the use case for such data. The patent classification system may work well for identifying functional characteristics of a focal patent, but the results suggest that it does not work well for identifying near-neighbors to a focal patent. In other words, a categorical assignment may be helpful for the USPTO when assigning a patent application to an art unit for examination, but be problematic for scholars when examining problems that are relational.

## Conclusion

A precise measure of technological space is important for many core questions in economics, strategy, and innovation. We developed and validated a new measure of technological space, and the underlying data is already supporting a range of additional research endeavors: Kuhn and Younge (2016) use similarity data to show how patent citation patterns have become dramatically noisier over time; Younge and Rmer (2016) use similarity data to control for the pre-existing distribution of inventive activity and contribute new evidence to the debate on the geographic localization of knowledge spillovers (Jaffe, Trajtenberg et al. 1993, Thompson and Fox-Kean 2005); and Thompson and Kuhn (2016) use similarity data to evaluate the effect of winning a patent race on cumulative innovation.

Patent similarity can be used to position firms (see, for example, Stuart and Podolny 1996): scholars can compare the countervailing effects of spillovers in technology space and business stealing in the product market space (Bloom, Schankerman et al. 2013); market entry can be detected by identifying cases when a firm first files a patent similar to portfolios by incumbents (Hall and Ziedonis 2001); and patent portfolios of allied (or competing) firms can be compared by patent similarity to measure the extent of overlap or divergence between firms. (Mowery, Oxley et al. 1996). Patent similarity can also be used to measure innovation: researchers can compare the similarity of a focal patent to patents issued before/after it to measure breakthroughs (Ahuja and Lampert 2001); and patent similarity can be used to measure the effects of venture capital funding by observing firms' technological

contributions (Kortum and Lerner 2000). Patent similarity can also be used to investigate the patent system itself: patent similarity can be used to identify patent thickets and/or holdup by investigating whether dense webs of patents deter product development (Lemley and Shapiro 2006); patent similarity can be used to analyze patent scope by comparing the effects of boundary-spanning patents on follow-on innovation (Merges and Nelson 1990); and the effect of winning or losing a patent race can be investigated by comparing outcomes by highly similar patents by different firms (Lerner 1997). The VSM similarity measure may shed new light on selection effects that are inherent to many applications of patent data: patent similarity can be used to construct robust counterfactuals for investigating phenomena such as geographic spillovers (Jaffe, Trajtenberg et al. 1993, Thompson and Fox-Kean 2005), and patent-to-patent similarity can also be used to better understand patent citations (Trajtenberg 1990).

Scholars have argued that existing measures of technological space are too coarse for many research questions (Thompson and Fox-Kean 2005), and we provide descriptive evidence in support of that argument. In addition, we develop a new measure of technological space and validates it for future research. Although classification-based measures remain useful for some uses, textual similarity is a better measure for researchers to use for the following reasons.

First, the VSM measure generates a single, standardized value that represents the same concept, even in different contexts and across different samples. It is conceptually simpler than patent classification, as there is only one number, it works the same for any comparison, and there are fewer (arbitrary) decisions to be made about the correct primary/secondary, class/subclass, and classification scheme to follow. Nevertheless, researchers also can use patent-to-patent similarity at a low-level to construct derivative measures at a higher level to meet specific research needs. Second, the VSM measure is continuous, bounded zero to one, free of breakpoints and divisions, and calculated in a way that is consistent across domains. Although measures based on a patent classification scheme may sometimes appear to be continuous (due to aggregation), classification itself is always discrete, for a patent either belongs to a class or it does not. Third, the VSM measure is free of selection effects and errors made during classification, as well as the lack of observations associated with many patent subclasses. Fourth, we demonstrate the validity of patent-to-patent similarity using a range of techniques. The measure is based on recent theoretical and empirical work in information retrieval theory. It correlates well with inventor, patent attorney, and lay person notions of relatedness; it predicts patent citations better than existing class-based measures; and it matches with patent phenomena both internal (patent families) and external (firm technological focus) to the patent system.

# Chapter 3

### Does Winning a Patent Race Lead to More Follow-on Innovation?

## Introduction

On February 14, 1876 two men filed patents for the telephone. One, Alexander Graham Bell, received the patent and started a company that was so successful that even today his name is synonymous with the telephone. The other, Elisha Gray, is usually only known, if at all, as the telephone inventor who wasn't Bell.

Their case is perhaps the most famous example of a "patent race"—where two entities race to invent and then get a legal monopoly for their invention. This situation has been much theorized about by, for example, Loury (1979), Lee and Wilde (1980), and many, many others. Typically this is modeled as racing firms making investments over time to develop an idea, with the one that develops it first getting patent coverage and therefore all of the profits from it.

Individually such work produces clear economic predictions, but collectively it produces strongly divergent predictions—with results that depend heavily on initial conditions, market power, information availability, etc. For example, some models predict that patent races will end almost as soon as they begin, because an initial lead by one entity will result in the laggard exiting the race. (Gilbert and Newbery 1982). Others allow for sustained racing, even if one entity is in the lead. (Fudenberg, Gilbert, Stiglitz and Tirole 1983, Harris and Vickers 1985). More modern models (e.g., Judd, Schmedders and Yeltekin (2012)) are subtler about their predictions, highlighting the range of outcomes that could occur, based on differences in empirical parameters.

Results from the real world have the potential to help focus this debate, but, to the authors' knowledge, no publications have emerged that empirically investigate the determinants or outcomes of patent races, except in very narrow settings, because of the difficulty in observing patent races. Because, although it is easy to see the patent that issues as a result of winning a patent race, how could one look for the non-existent patent from the loser? And, without seeing both, how would one know that the winning patent was part of a patent race, and not just a typical instance of invention and patenting?

We introduce a novel strategy for observing patent races. It does not allow all patent races to be observed—in particular, it misses case where one competitor drops out very early in the race. If patent races followed the pattern of the early, deterministic theory literature, this would be a problem, as those models predicted that, as soon as one competitor took a slight lead, the other would drop out of the race. (Lee and Wilde 1980, Loury 1979). Under those circumstances, our method for detecting patent races would observe no racing. In practice, this method shows that there are many patent races that occur well into the development stage, when they are observed. In this, these findings are supportive of models where firms accumulate knowledge that can be used towards today's patent race, but which also helps towards future R&D (including other races). (Doraszelski 2003, Fudenberg, Gilbert, Stiglitz and Tirole 1983, Harris and Vickers 1985, Harris and Vickers 1985).

We employ a method for observing patent races that takes advantage of data that can be gathered from the patent system and the empirical observation that often in patent races,

both competing firms will apply for a patent on the contested discovery, but only one will get it. The reason that both will apply is that there is a gap in timing between when a patent is filed and when it is made public. During this window a patent has "priority"—future patents that claim the same invention will be disallowed—but it is not yet public knowledge. Thus, during this window, two patent applications can be filed with the patent office for the same invention with both applicants believing that they are first, but only the first applicant will receive the patent coverage whereas the second will have its claims rejected. Importantly, however, both applications will be present in the patent office's records. Using this administrative data allows the observation of the presence of these events and the construction of "patent twins"—sets of patent applications filed at nearly the same time on the same invention. These patent twins are then contrasted with all other patents to observe where and to what extent patent racing is happening within the patent system. The outcomes for the winner of the patent race (the "leader") are also contrasted with the loser of the patent race (the "follower") to observe the effects of winning the patent race.

Although this method misses any 'over before they start' races—it nevertheless shows that patent racing is a frequent occurrence, with 10-11% of all patents being in a race. Moreover, we provide strong evidence at the project level that winning a patent race is valuable, leading the winning firm to maintain their patent coverage more often.

We also consider the causal effect of winning a patent race on follow-on research. Despite the large sample, the data reveals no evidence of firm-level differences in spending on research and development or in profitability from winning an individual patent race. Moreover, one should not expect to see such a difference, because the diverse research (and patenting) portfolios of large firms lead them to both win and lose large numbers of patent races, and thus the net number of 'wins' should be small. Indeed, the data shows that the largest firms win and lose patent races in roughly even proportions.

However, when considering the causal effect of winning a patent race at the project level, we show that firms that win a patent race are 11% more likely to do related follow-on research, and that the total volume of related work done by them is 14% greater than that done by those that lose the patent race. Finally, it shows that the follow-on work done by patent race winners is more closely related to the work covered by their patent, suggesting that they are exploiting their patent-protection while firms that lose the patent race are inventing around.

This chapter proceed as follows. First we introduce both the theory (Section 3.2) and reality (Section 3.3) of patent racing, and show that there are large and consequential differences between them. In Section 3.4, we discuss where patent races happen and which firms participate. In Section 3.5 we discuss how we get a causal estimate of the effect of winning a patent race. Section 3.6 presents these causal estimates, at the patent, firm, and project level. In Section 3.7 we conclude.

### Patent Races in Theory

Patent racing occurs when firms compete to produce and patent a new invention first. Because the patent system provides a legal monopoly over the technology to the winner, winning a patent race allows a firm to extract above-normal profits by selling the invention in a market where competitors are unable to replicate it. Thus, models of patent racing tend

to treat a patent as a winner-takes-all prize that is awarded to one of a continuum of racing firms that discovers an independent, unitary invention. Early models included processes for discovery that were either (i) random, or (ii) deterministic. In early random-discovery models, firms compete vigorously until a patent is obtained. (Dasgupta and Stiglitz 1980, Lee and Wilde 1980, Loury 1979, Reinganum 1981, Reinganum 1982). However, a firm could never pull ahead of its rivals because a firm's past R&D was assumed not to affect its current likelihood of discovering an invention. Instead, firm strategies in equilibrium induce a welfare tradeoff between the benefits of faster innovation and the costs of overinvestment in R&D, yielding the result that firms invest in R&D beyond the socially optimum level when market structure is fixed. (Loury 1979), (Lee and Wilde 1980). In early deterministic-discovery models, each firm made observable progress toward the prize. However, the equilibria of these models was $\epsilon$-competition, unobserved in the real world, in which a firm with an arbitrarily small head start causes all (identical) rivals to immediately drop out of the race because they have no chance of winning. (Dasgupta and Stiglitz 1980, Gilbert and Newbery 1982).

Subsequent articles identified conditions under which firms could leap-frog each other in making progress. A multi-stage R&D process with random discovery can support leapfrogging under perfect competition because a firm running behind can catch up with enough time and luck, and so they will attempt to do so if the expected return winning the patent race is large enough. (Fudenberg, Gilbert, Stiglitz and Tirole 1983). Leapfrogging can also occur with deterministic discovery if firms differ in terms of their "(1) their valuations of the prize, (ii), their discount rates, (iii) their efficiency at making progress, and (iv) their initial distances from the finishing line." (Harris and Vickers 1985). Finally, leapfrogging can occur if firms accumulate knowledge over time and that accumulation reduces the cost of future R&D. (Doraszelski 2003).

The findings presented in this article show that patent races are common, and that many proceed well into the development stage of the idea. Moreover, these findings provide direct evidence that knowledge accumulated during the patent race typically provides advantages to both the winner and the loser of a patent race.

### Patent races in the wild

The empirical work on patent races is thus far extremely limited. An early study of the disk drive industry finds support for Reinganum's patent racing model. (Lerner 1997). Later, a laboratory experiment with 36 subjects provided limited support for the Harris and Vickers multi-stage model. (Zizzo 2002). This article extends the empirical literature by being the first study to patent races throughout the patent system, the first to catalogue where they occur, and the first to causally identify the effects of winning or losing a patent race using observational data. To do this, we start by describing institutional features of how patent racing happens in the wild.

### Continuous Technology

In a canonical patent race, firms compete to patent a discrete idea. In reality, technology is more continuous. Even if two firms work towards the same discovery, the specifics of their implementation are likely to differ because of design and development choices. These

could be subtle differences (e.g., different programming languages used to develop the same functionality) or substantive differences in otherwise similar inventions (e.g., chairs made of different materials). The latter case resembles models of knowledge accumulation during patent racing, in that a firm might not have been the first to discover a property of chairs, in general, but might nevertheless generate an R&D advantage on doing so with a particular type of material.

Such differences in technology also manifest themselves in the patenting of technology. Two firms may try to claim a common idea, but might also claim ideas that are not shared with the other. For example, if firm 1 claims a set of ideas, A, first while racing with firm 2. If firm 2 claims a set of ideas, B, then firm 1 should 'win the race' on A∩B, and uncontestedly win the remaining claims in A, whereas firm 1 'loses the race' on A∩B, but still uncontestedly wins the non-intersecting claims in B.

Depending on the size of firm 2's non-intersecting claims they should be likely to abandon the application (if no claims remain, or there is too little value in those that remain to justify the filing or legal costs) or pursue the patent to issuance but with a narrower scope (if the non-overlapping claims still have value).

### Winning a Patent Race

Patent races also diverge from theory in the mechanics of how they are won. Instead of a universally-acknowledged victory of a known benefit, there is typically no announcement of victory, and trailing firms may continue to race even after the race has been won. Moreover, even once the victor is known, the spoils of that victory (patent protection) are often unclear until years afterwards.

The USPTO makes the existence and content of a patent application public to the world no sooner than 18 months after its earliest priority date, which is typically its filing date. This delay creates a window of time in which a competing firm could continue racing, unaware that it had already lost the race. Such firms may then try to claim patent protection for ideas that have already been submitted to the patent office by the earlier-filing firm. In such cases, the patent office will ultimately reject the claims that duplicate coverage.

The determination of whether a patent gets issued at all, and if so what claim scope is approved, typically takes 3 years to complete. (USPTO 2015). As such, it is only at this point that patent racers know with some certainty the value of the patent protection prize that they have won. (See, e.g., Gans, Hsu and Stern (2008)).

Because patent races differ from theory in having these ambiguous endings, it is common for two patent racers to file for patent protection on the same idea within a short period of time, in a more-modern, firm-version of what Alexander Graham Bell and Elisha Gray did 140 years ago. The administrative records from these pairs of applications are the empirical window through which patent racing is observed for the purpose of this analysis.

### How the patent system lets us observe patent races

Patent examination is an extremely complex and frequently lengthy process of procedure and negotiation. Both the patent applicant's choices and each of a series of decisions reached at the patent office exert considerable influence on not just whether a patent application is

ultimately granted, but also the time required to come to a final decision as well as the scope and strength of the claims of a patent if one is issued.

At a high level, the lifecycle of a patent application is divided into pre-examination, examination, and issuance. During pre-examination, the patent application is classified by technology area and assigned to a patent examiner knowledgeable in the relevant field. The examination phase begins when the patent examiner has time to begin work on a new application. During the examination phase, the patent examiner evaluates the application to ensure that it complies with every requirement of U.S. patent law. The USPTO is required by U.S. law to grant an application for a patent unless it can find a reason to reject the application, but the requirements imposed on patent applicants are numerous and complex. Reasons for a rejection can range from the administrative (e.g., a patent applicant has made a procedural mistake in the examination process) to the descriptive (e.g., a patent application is improperly formatted) to the substantive (e.g., a patent application is insufficiently inventive).

This analysis focuses on substantive rejections. In particular, on novelty rejections - the case when an examiner rejects one or more of the claims in a patent application on the grounds that someone has already disclosed it to world. Under U.S. law, a patent claim is anticipated (i.e. is not novel and therefore not patentable) if a single prior art reference discloses all of the features recited in the claim. Federal courts have interpreted the novelty requirement as meaning that "[a] claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference." Further, U.S. law requires that "[t]he identical invention must be shown in as complete detail as is contained in the ... claim" in order to support a novelty rejection. Thus, a novelty rejection specifically highlights the appearance of an invention that was earlier disclosed, but which a later-filed application has tried to claim. The following example helps to clarify the nature of novelty rejections.

In September 1980 Daniel and Vicky Lyer filed an application for "a wearer adjustable article carrying harness assembly adapted to releasably hold an article, such as a camera, against the body of the wearer in a carrying position while not in use ...". In March 1982 this patent, 4,320,863 was granted.

In October 2005, Nicholas Woodman, the founder and CEO of the action camera manufacturer GoPro, filed an application for "a configurable wrist or arm worn camera harness ... that facilitates convenient carry, access, and secure use of a camera during participation in a physical activity". In that application, Woodman made 47 claims for patent protection to cover his invention. However, the patent examiner rejected much of Woodman's claim coverage for lack of novelty, saying "Claims 1-3, 5-12 and 30-45 are rejected under 35 U.S.C. 102(b) as being clearly anticipated by Lyer et al." As a result, Woodman dropped all these claims. Nevertheless, Woodman still ended up with patent 6,955,484 which covered the 17 claims that remained after the novelty (and other) rejections.

This example highlights that patent applications with rejected claims can still produce an issued patent and, in fact, this is the norm for patent races. This example also shows some of the standard phrases that can be used to identify novelty rejections ("rejected under 35 U.S.C. 102(b)"). Novelty rejections are analytically useful for us for several reasons. Firstly, the examiner indicates the earlier disclosure which described the invention. When that disclosure is a patent, this provides a direct indication that the examiner believes both

patents cover the same invention. Under the right conditions, such a rejection can be indicative of a patent race. Moreover, the existence of the potential patent race was determined by a third-party expert evaluating the claims contemporaneously as part of their job. This determination saves us, as econometricians, from having to estimate the existence of the race from indirect measures or from after-the-fact retrospection about technological similarity.

Importantly for the identification strategy, a patent applicant who receives a novelty rejection only has a limited set of options: they can abandon the application, argue that a rejection is improper, or amend it. Those that amend cannot add new matter to the application—the application's inventive contribution is fixed at the time of filing. However, they can narrow their claims, so that they don't overlap with the earlier work, and pursue the issuance of this more-modest patent. Importantly, applicants face disincentives from filing exceptionally broad patent applications since doing so is unlikely to result in a broader patent and will be more costly due to the extra time and effort required to issue the patent.

### Patent priority

Thus far, the analysis has assumed that patent coverage is granted to the first party to file that discovery with the patent office. In reality, patent coverage for patent applications filed prior to March 16, 2013 was ostensibly granted to the first party to invent rather than the first party to file. A party that submitted a patent application for which the patent examiner found prior art could submit documentary evidence that demonstrated that the actual invention predated the time of filing, thus "swearing behind" the prior art. However, applicants attempted to swear behind the filing date of only .6% of patents issued between 2000 and 2005, and the percentage is even lower outside of biotechnology. (Crouch 2010).

If two patent applications (or one patent application and an issued patent) claim essentially the same invention and have conflicting priority claims, then the USPTO can declare an interference proceeding to resolve the dispute. In practice, interference proceedings are uncommon due to the high cost associated with pursuing them, which typically exceeds $650,000 in attorney fees. According to one practitioner, "Less than one percent of the applications filed become involved in an interference proceeding, and in fiscal year 2007 only 7 interference proceedings resulted in the patent going to the second to file, who was determined to be the first inventor." Quinn (2010). Thus, even though in theory the patent system was first-to-invent before 2013, in practice it has been overwhelmingly first-to-file throughout the sample period, and hence this analysis assumes that the first patent to be filed gets any contested claims.

### Data

When a patent examiner evaluates a patent application, they prepare various documents and post them to the USPTO's Patent Application Information Retrieval (PAIR) system. For patent applications that have not yet been published, these documents are accessible only to the patent applicant and the applicant's designated representatives. All documents generated during the examination of the application become publicly accessible via the same system when an application is published as a patent publication or issued patent, typically 18 months after filing. Documents for patent applications that are abandoned before being

published remain private forever, and data is limited even for applications that are published prior to abandonment, which severely limits the analysis that can be done on abandoned patents.

Thus, most of the analysis focuses on instances where both the leader and the follower issued as patents, which in fact make up the majority of all patent races observable in the data. This has the potential to introduce a selection bias in the estimates for the following reason: Presumably applicants abandon poorer quality patents and patents that lose a patent race are likely to be over-represented in this group since, by definition, they lost patent coverage compared to the corresponding patents that won. As a result, insomuch as issued patents that won are observed to be of higher quality, this will understate the size of the effect that would have been observed had the full distribution of losing patents been observed. The estimates are thus biased towards zero, and thus the coefficients should be considered lower bounds.

Through an agreement with the USPTO, Google scraped the PAIR system and aggregated publicly available patent examination documents in ZIP files that are accessible through Google's website. The full data includes PAIR data for 5,027,882 patent applications filed between April 23, 1984, and June 10, 2013. Data availability considerations restrict this further to a sample of patents issued between January 1, 1998 and February 7, 2012.

**Table 4: Summary statistics.**

| | |
|---|---|
| Patents | 5,027,782 |
| Earliest filing date | April 23, 1984 |
| Latest filing date | December 11, 2014 |
| Claim rejections | 18,335,938 |
| Novelty rejections | 1,265,038 |
| Mean specification length (words) | 9,274 |
| Mean number of claims | 17.11 |
| Mean number of inventors | 2.592 |

We extracted from the PAIR data 4,231,280 Office Action documents. Office Actions are published in a PDF image format that renders their text inaccessible in a digital format. Accordingly, we first applied the optical character recognition library Tesseract to recover the text of Office Actions in a digital format. Because the application of the Tesseract library to so many documents required many years of computation time, we divided the computation among 10,000 CPUs accessed via the Amazon Elastic Compute Cloud (Amazon EC2) service.

We searched the text of each patent application's Office Actions to identify patent rejections. Patent rejections are not listed in an Office Actions in an easily accessible format or location, but rather are embedded throughout the text of the document. When stating claim rejections, however, patent examiners employ formal legal language to clearly indicate the legal status of each claim. For example, a patent examiner issuing a novelty rejection ("35 U.S.C. 102") would include a sentence such as "Claims 1-20 are rejected under 35 U.S.C. 102 as anticipated by U.S. Patent No. 1,234,567 by Smith et al." In the paragraphs that follow, the patent examiner would support the rejection with documentary evidence comparing the claims of the application to the disclosure of Smith. We applied regular expressions to

the text of each Office Action to identify rejection sentences, determine the statute used to support the rejection, and extract the patent or patent publication number of the reference used to support the rejection.

We linked each patent in the dataset with bibliographic data collected from other publicly available documents issued by the USPTO. Each week the USPTO issues an XML document that include a wealth of tabularized data describing each patent application published or issued during the week. From this data we extracted such fields as the patent filing date, the patent issue date, and all patent citations made by each issued patent.

The sample is limited to those pairs where neither patent application claims priority to a previously-filed application to avoid evaluating continuations or first-foreign-filed patents as potential patent races, both of which could imply disclosure patterns different than the 18-month window.

Each issued patent is also linked with patent maintenance fee payment data parsed from the USPTO's maintenance fee payment data file. To maintain a patent in force, the patentee must pay a fee to the USPTO by each of 3.5, 7.5, and 11.5 years following the issue date of the patent. The maintenance fee payment data file lists for every patent whether or not the patentee made each of these payments. Maintenance fee payment data can provide some indication as to the private value of a patent, as a patentee would only pay a maintenance fee if the patent were more valuable than the fee amount.

We retrieved information on publicly traded firms via the Compustat database. Linkages between patents and firms were provided by Douglas Hanley, who parsed USPTO patent applicant data to associate each patent filed by an institution with an institution identifier, and who linked patents issued to publicly traded companies with a Compustat identifier. (Hanley 2015). Firms that are not present in Compustat are dropped from any sample that considers firm-level characteristics.

### Defining Patent Races

To support the analysis into the effects of winning or losing a patent race, the notion of a "patent twin" is introduced. By analogy to Bikard's 'paper twins,' a patent twin refers to two patents filed at roughly the same time, covering the same discovery. (Bikard 2015). We first present an example, and then codify this definition.

In April 2005, Cardiac Pacemakers Inc. filed U.S. Patent No. 7,640,057, with the title: "Methods of providing neural markers for sensed autonomic nervous system activity." Eight months later, Pacesetter Inc. filed U.S. Patent No. 7,672,722, with title: "Hardware-based state machine for use in discriminating near field signals from far field signals for use in an implantable cardiac stimulation device." Both applications sought protection for techniques for identifying cardiac events, as well as other aspects of their discoveries. Ultimately both were granted patents, but some of the scope that Pacesetter Inc.'s wanted was refused by the patent examiner on the basis of the earlier patent submission by Cardiac Pacemakers.

We define patent twins as pairs of patent applications that exhibit two characteristics. First, the two applications include sufficient technological overlap to claim patent protection for the same invention. In practice this is operationalized by looking for cases when an examiner has rejected claims by the later-filed application (which are termed the "follower" in this analysis) on the basis of those claims already being present in an earlier-filed one

(which are called the "leader" in this analysis)—that is, through novelty rejections. Second, the difference in time between the filing of the leader and the follower is sufficiently small that it would be unlikely for the applicants to know the content (and likely even the existence) of the others' patent application. In practice this analysis employs two time windows:

"All Twins 18": Applications within 18 months of each other

"All Twins 6": Applications within 6 months of each other

The 18-month window reflects time between when the USPTO would receive and make known to the world the contents of the leader application. We find this to be the most-compelling high-level definition for a Twin. It should be noted that a small percentage of patents take part in more than one patent race, such as when three firms all file competing applications.

In Section 3.5, however, infers a causal interpretation to the estimates for winning a patent race by arguing (and testing) that with a small enough time window, the difference between the 'leader' and 'follower' in a patent twin is as good as random. That is, Section 3.5 makes a regression discontinuity argument. For this, the analysis employs as narrow a bandwidth as possible. The 6-month window was chosen because it shows excellent covariate balance between leaders and followers, and still provides a sufficient sample size for the analysis.

In order to look at firm-level outcomes, there will also be instances where the sample is restricted to observations where both twin patents were filed by public firms. To parallel the definitions above these will be referred to as "Compustat 18" and "Compustat 6."

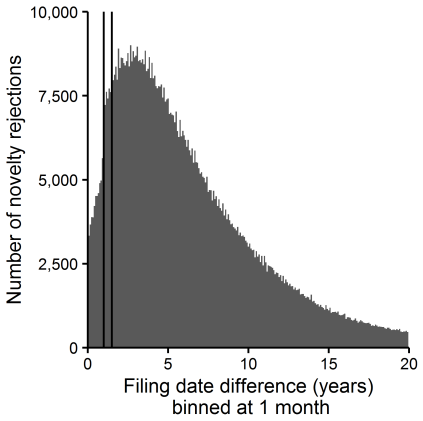



**Figure 8: Histogram of rejections by difference in filing date.**

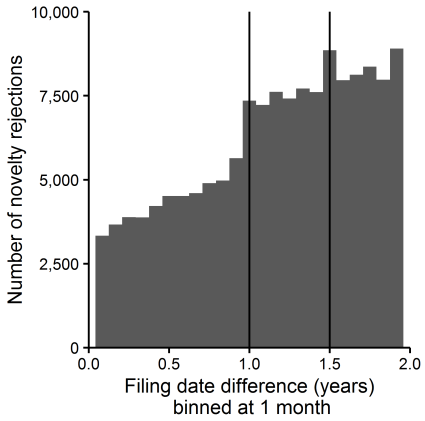



**Figure 9: A continuation of Figure 8, with a detailed view of years 0-2.**

The set of patent applications linked by novelty rejections is large and includes many instances where the filings are years apart. Such a situation might occur, for instance, if the follower is simply at a much earlier stage of technological development than the leader and is unaware of the state of the prior art, or could not afford to do a prior art search. We do not regard such situations as patent races because there is no active competition for the idea—one has already completed the race, and that information is public knowledge

prior to the other filing a patent application. To identify actual competition, the analysis is narrowed to application pairs, linked by novelty rejections, with "sufficiently small" time between their filing dates. Figure 8 shows the distribution of these differences in dates. Figure 9 is a sub-figure that focuses on the 0-2 year window after filing.

For the description of the phenomenon of patent racing, 'sufficiently small' is defined as the window of secrecy provided to patent applicants under U.S. law. Each patent application is secret upon filing. However, by default, every application filed on or after January 1, 2001 is made public 18 months after filing, so an applicant for a later-filed application cannot search for that earlier-filed application if it has been filed less than 18 months prior. The definition for a patent race used in this article is therefore: two firms filing patents covering the same invention within a window where the application of the other would not have been publicly known. Later, to do a regression discontinuity estimation, the window is narrowed even further to 6 months. However, the remainder of this section focuses on the 18-month window.

This methodology purports to find races for a common technology. Under such circumstances we should expect a great deal of similarity between the patents (although not perfect similarity, because each will also contain non-overlapping technology). We can test the success at this by comparing the textual similarity of technology descriptions for patent twins as compared with a reference group (say, patents in the same class). If the method is working, the technology descriptions of patent twins should be much closer. To test this, the analysis employs the textual similarity measure developed by Younge and Kuhn (2015), described in Section 4, which treats each patent specification (description of the technology) as a bag-of-words, and looks for similarity in the usage of non-common words. Figure 10 shows the distribution of these similarity scores.

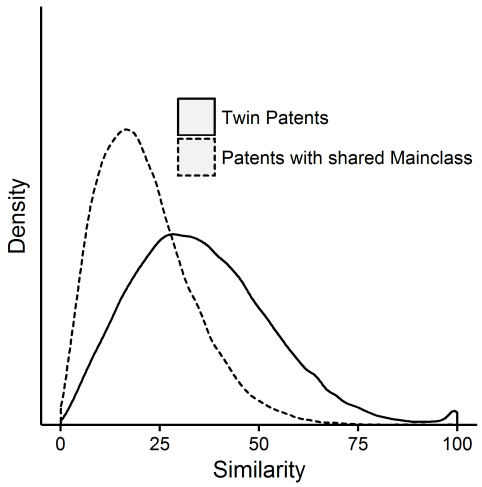



**Figure 10: Similarity (as measured by cosine distance) between patents in the same (main) patent class or patent twins. Scores only calculated for cases with similarity greater than a (low) minimum threshold for computational reasons.**

The results show that, as predicted, patent twins are much more similar than are patents in the same main class. In addition to being relatively more similar, they are also very similar

in absolute terms. For reference, less than 3% of patent pairs have a similarity score greater than 0%, and only .54% have a similarity score of 25% or greater.

Thus the method for identifying patent races, in addition to being based on criteria that mean it should correspond to patent races, can be shown to identify instances of patenting that are both close in time and close in the technologies where patent protection is sought. Therefore, this definition of patent twins allows the observation of patent racing.

### Where do Patent Races Happen? Who races?

This section shows that 10-11% of patents are in patent races. It also describes features of those races, where these races occur, and which firms do the racing.

### Description of Patent Races

Faced with diminished potential patent coverage because of a lost patent race, a firm can choose to either abandon its application or pursue issuance on the reduced scope. In practice, abandonment is the rarer outcome. More common is for firm 2 to pursue, and get, a narrower patent. In the observed races, only 33% of those that lose a patent race abandon their patent (compared to 12% for those that win). The remaining 67% of firms that lose a patent race still end up with a patent, but that patent will have narrower scope. That is, they will have protection over a smaller area in 'technology space,' or their protection in an area will be less strong.

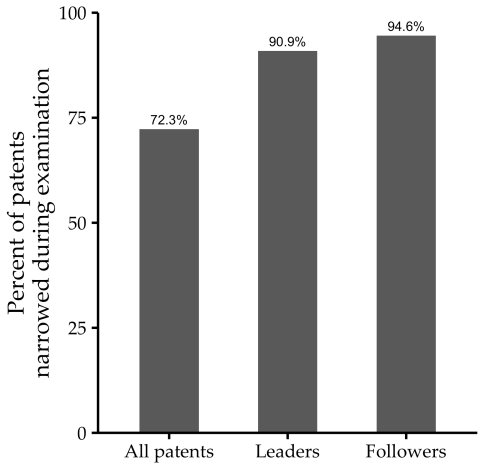



**Figure 11: Percentage of patents narrowed during examination at the USPTO.**

In a separate article, the authors find that 70% of patents from a random sample have their scope narrowed during prosecution (as judged by patent lawyers and new metrics). (Kuhn and Thompson 2017). This same metric is applied in this article to look at patent scope narrowing in patent races. Because patent races are (by definition) a sign of a contested technology area, and thus of a high amount of potential conflict in patent claims, it is not surprising that above-average patent narrowing occurs for both the winning and losing patent, and that those that lose a patent race are particularly hard hit. Figure 11 shows that

this is true. Whereas 72% of all patents get narrowed during the prosecution process, 91% of winning patents ('leaders') and 95% of losing patents ('followers') get narrowed.

The greater narrowing of patent twins suggests that patent races happen disproportionately in areas crowded with competing patent claims. This might be driven by the market structure in these areas, or perhaps by the type of technology. More interesting for theory, however, is that it might indicate that patents in races reflect smaller innovative steps. If so, they would be 'closer' to existing patents and thus would be more likely to be narrowed by them.

### Technology Areas

Figure 12, Figure 13, and Figure 14 show the technology areas where patent races occur. Here, patents are divided by their classification into USPTO technology centers. Figure 12 shows, in absolute numbers across all the years of our data, that the most patent races happen in the semiconductor and communications divisions, and fewest happen in the biotechnology division. However, Figure 13 shows that, as a share of all patents in those areas, Computers has the largest share (15.7%), with Communications next (13%), and then Semiconductors (10.1%). Biotechnology also has the smallest share, with 4.9%. Finally, Figure 14 recapitulates Figure 13, but instead of reporting the share of all patents in that technology center, it considers those in patent races as a share of those that had novelty rejections (at any point in time). This confirms that it is not just a prevalence of novelty rejections in these areas that are driving these results, but instead a higher rate of contemporaneous patenting of the same technologies—that is, more patent racing.

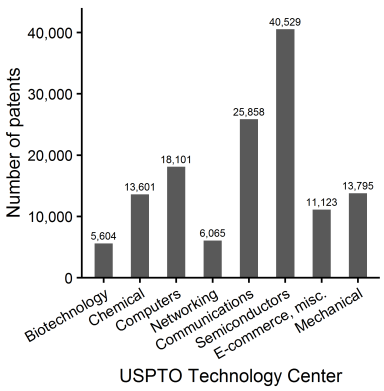



**Figure 12: Patents in patent races.**

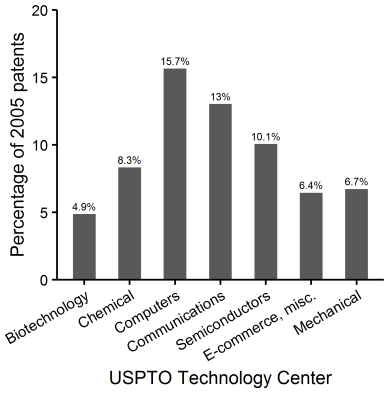



**Figure 13: The percentage of patents issued in 2005 that are in at least one twin, binned by USPTO technology center. The networking technology center is omitted because it was started in 2005 and not yet issuing patents.**

The reason behind these differences is unclear, and calls for more research. Plausible explanations could include (i) technology lifecycles, (ii) patent effectiveness, or (iii) market

structure. These patterns could reflect technology lifecycles if areas like computers and communications require smaller inventive steps to get patent coverage. If so, these could lead to a greater density of patents at the frontier of knowledge and more competition (i.e. patent races) for the open spaces.

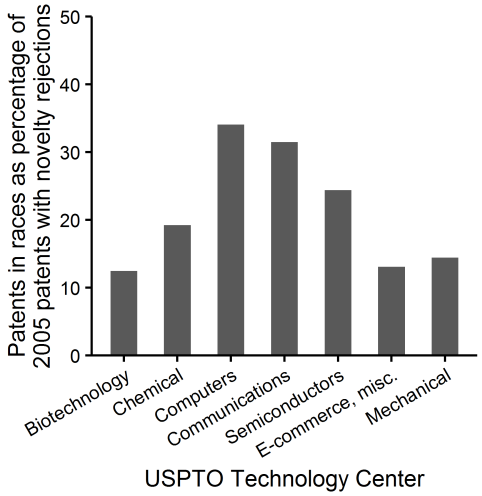



**Figure 14: Patents in races, as a percentage of patents with novelty rejections (2005).**

Alternatively, differences in patent effectiveness could cause this distribution of patent races. Biotechnology and Mechanical are technology areas thought to have a stronger ability to enforce patents, whereas Computers and Communications are thought of as having a weaker ability to do so. (See, e.g., Cohen, Nelson and Walsh (2000)). If firms deem themselves to have a significant chance of losing the race, they might choose not to race in the first place if they think that they would have little ability to continue the work in the face of a strong competitor patent.

Lastly, the distribution of patent races might reflect market structure. For instance, there could be fewer patent races in some areas simply because there are fewer (or less direct) competitors. This could even be an endogenous result of the cost of patent racing if, in areas like pharmaceuticals, it is so costly that it has led to consolidation in industry structure.

### Firms

How are firms that patent race different from other firms that patent? To answer these questions, the data is narrowed to publicly-traded companies filing applications within 18-months of each other (that is, to our "Compustat 18" dataset).

The restriction to public companies also implies narrowing our sample to instances where both the applications of the leader and the follower issue as patents because the USPTO does not require applicants to provide the relevant information needed to match patents to firms until patent issuance. As a result of this data constraint, this analysis misses patent races where the follower abandons their application. However, high value patents are unlikely to be abandoned (Kuhn and Thompson 2017), so any selection introduced is likely to affect mostly low-value patents.

Figure 15 illustrates the share of our data in these technology centers for each of these samples, again using 2005 as the representative year. It shows that narrowing to public-firm patent races leads to a slight over-representation of computers and communications and a slight under-representation of biotechnology, e-commerce, and mechanical, as compared with the full sample.

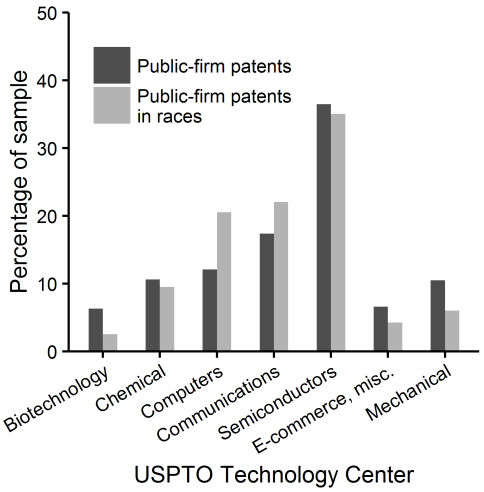



**Figure 15: Sample Composition—Public-firm patents in races compared to all public-firm patents (2005).**

Table 5 and Table 6 compare firm characteristics for those associated with twin patents to those associated with patents in general. Not surprisingly, firms in patent races are larger and do more R&D, a difference that exists even on a per employee basis, where racing firms have slightly higher sales and R&D per employee. As Table 7 shows, they are also much more prolific patenters, both in terms of filed and issued patents.

**Table 5: Comparison of the firm characteristics of twin patents to firm characteristics of all patents. Full sample 2005.**

| Variable | Mean public | Mean twins | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Revenue (twin filing) | 8.918 | 8.626 | 3.762 | 0.0002 | 876 | 0.140 | 0.444 | 0.292 |
| RD expenditure (twin filing) | 6.521 | 6.557 | −0.617 | 0.537 | 873 | −0.153 | 0.080 | 0.037 |
| Assets (twin filing) | 9.276 | 9.064 | 3.133 | 0.002 | 881 | 0.079 | 0.345 | 0.212 |
| Employees (twin filing) | 3.588 | 3.325 | 4.753 | 0.00000 | 860 | 0.154 | 0.372 | 0.263 |
| Cost of goods sold (twin filing) | 8.306 | 7.912 | 5.031 | 0.00000 | 876 | 0.240 | 0.548 | 0.394 |

**Table 6: Comparison of the firm characteristics of twin patents to firm characteristics of all patents, on a per-employee basis**

| Variable | Mean public | Mean twin | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Revenue / employee | 298.0 | 310.9 | −3.581 | 0.0003 | 2,605 | −19.867 | −5.809 | 12.838 |
| RD / employee | 37.7 | 41.0 | −3.208 | 0.001 | 2,597 | −5.377 | −1.298 | 3.337 |
| Assets / employee | 474.0 | 462.7 | 1.324 | 0.185 | 2,544 | −5.410 | 27.926 | 11.258 |
| Cost of goods sold / employee | 170.6 | 165.5 | 2.109 | 0.035 | 2,722 | 0.358 | 9.852 | 5.105 |
| Employees (logged) | 3.6 | 3.8 | −6.281 | 0 | 2,482 | −0.287 | −0.150 | 0.218 |

**Table 7: Comparison of the (logged) firm characteristics for twin patents to firm characteristics for all patents. Full sample 2005.**

| Variable | Mean all | Mean twin | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Prior apps. filed (2 yrs) | 4.434 | 6.639 | −55.194 | 0 | 2,425 | −2.283 | −2.126 | 2.205 |
| Prior apps. filed (5 yrs) | 4.991 | 7.280 | −53.638 | 0 | 2,426 | −2.373 | −2.205 | 2.289 |
| Prior pats. issued (2 yrs) | 3.701 | 5.534 | −31.013 | 0 | 2,336 | −1.950 | −1.718 | 1.834 |
| Prior pats. issued (5 yrs) | 4.173 | 6.111 | −30.350 | 0 | 2,337 | −2.063 | −1.812 | 1.938 |

## Methodology for Obtaining a Causal Estimate of the Effect of Winning a Patent Race

### Regression Discontinuity Approach

As previously outlined, the firm-level sample considered so far has been "Compustat 18," which represents patents linked by a novelty rejection that were filed within 18 months of each other. Within this 18-month window firms may race for patent coverage without having the earliest-filed patent application revealed. That does not mean, however, that within this window firms are the same. One might imagine that patenting leaders might be systematically different than patenting followers, for example they might do more R&D or work with better lawyers. And, indeed, tests of covariate balance at the 18-month window show a lack of balance at this point.

This creates a challenge for one of the empirical goals, estimating the causal effect of winning a patent race, because it indicates selection. To address this concern, we adopt a regression discontinuity approach. Imagine that two patents covering the same invention get filed at the patent office on the same day. Absent intrusive (and possibly illegal) corporate espionage, the exact timing for each application on that day is likely to be as-good-as-random. Under these conditions, small changes in filing time (the running variable) will introduce large changes in patent protection (the treatment)—and thus differences in firm outcomes should be interpreted as arising causally from winning the patent race. Figure 16 shows this schematically. Our estimation focuses on the effect of this loss of patent coverage, due to losing the patent race.

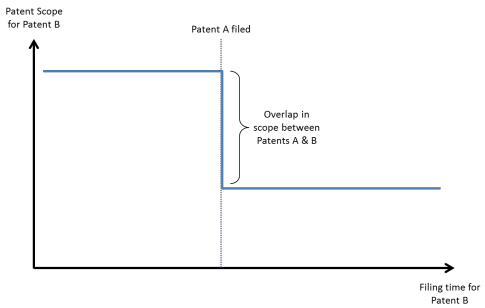



**Figure 16: Regression discontinuity setup.**

Unfortunately for the empiricist, there is a dearth of same-day, same-invention filings at the patent office, so a larger bandwidth must be used for the comparison. The larger the bandwidth we choose, the more data we get to consider, but the greater the chance that other factors confound the estimates.

The analysis employs a 6-month window as the bandwidth because this is a plausibly short window wherein small random events (e.g. illness of a patent attorney) could lead to changes in the order of filing. If this assertion is correct, that is, if filing order within these six months is as-good-as-random, we should see good covariate balance on pre-treatment variables between those filing patents that will ultimately win the patent race and those that will ultimately lose it. This is tested in several ways, as it is central to the causal interpretation of the effect of winning a patent race.

### Testing Whether those that Win and Lose Patent Races Look Similar, Pre-treatment

Testing for covariate balance on the full set of data is only possible in a limited way because some leader and follower patent applications are abandoned, and the USPTO publishes very little data on abandoned applications. As a result, the balance check on the full set of data focuses on the small number of covariates that are both available and of some interest.

**Table 8: Application-level balance.**

| Variable | Leaders | Followers | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Is small entity | 0.208 | 0.183 | 9.841 | 0 | 96,802 | 0.020 | 0.030 | 0.025 |
| Title Length | 61.829 | 60.984 | 4.216 | 0.00002 | 97,211 | 0.453 | 1.239 | 0.846 |
| Has Cust. No. | 0.876 | 0.884 | −3.913 | 0.0001 | 97,459 | −0.012 | −0.004 | 0.008 |
| Cust. No. Patents (log) | 7.976 | 7.858 | 9.144 | 0 | 86,126 | 0.093 | 0.144 | 0.119 |

Although all the differences in Table 8 are statistically significant, the significance is largely a product of the large sample size. In terms of economic importance, we observe limited differences. Slightly more patent leaders are small entities, 20.9% vs 18.5% for followers. The length of their patent titles are almost identical—61.8 vs 60.9 characters. The analysis also allows the observation of some characteristics about the entity that files the patent (which may be the firm, but is commonly their attorney) by looking at the USPTO

customer numbers associated with leader and follower patents. The results show that slightly fewer (87.5%) of leaders have USPTO customer numbers, vs 88.3% of followers, although their customer numbers are associated with slightly higher numbers of patents. Because you must have a USPTO number to file electronically, these differences probably reflect small changes in the timing of different attorneys going digital. To emphasize how similar leaders and followers are at the time of filing, Figure 17 and Figure 18 show the full distribution of title length and the number of patents associated with a customer number. Thus, we conclude that although there is evidence for small differences between the leaders and followers at the time of filing, the potential for these to bias the estimates is limited.

Because balance on the full set of data can only be looked at with a limited number (and relevance) of covariates, we also look for balance across a much larger set of variables, post-abandonment. Because, not surprisingly, followers are almost twice as likely to abandon their patent application (32.8%) as leaders (17.9%), the following tests are conditional on issuance, and the effect of differential abandonment rates must be kept in mind in interpreting the results. That said, since this effect would downward bias the estimates, the findings still provide a lower-bound for the effect sizes that result from winning a patent race even if such selection exists.

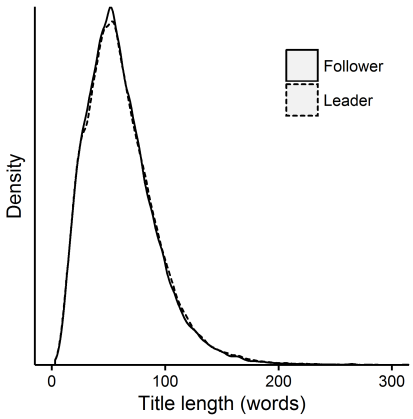



**Figure 17: Distribution of title length for leader and follower patents at the time of filing.**

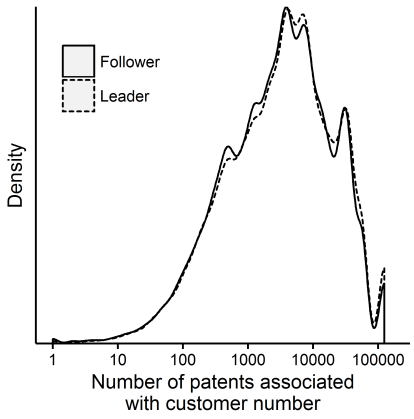



**Figure 18: Number of patents associated with the customer number (often the attorney) for leader and follower patents at the time of filing.**

Table 9 compares Student's T-test values for firm-level financial characteristics of leaders and followers in the sample. Because, by construction, follower patents are filed later than leader patents in the sample, we analyze firm variables as of the filing year of the leader for both.

The firms associated with leaders and followers look very similar, both in terms of the magnitudes of their characteristics and the lack of statistical significance between them. They exhibit no significant differences on revenue, assets, number of employees, or cost of goods sold as of the year in which the leader application was filed. R&D levels are statistically different at a threshold of 10%, but the magnitude of this effect is incredibly small.

Beyond firm financial performance, we also look for difference in previous patenting

**Table 9: Financial characteristics (logged).**

| Variable | Leaders | Followers | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Revenue | 9.280 | 9.357 | −1.310 | 0.190 | 4,490 | −0.192 | 0.038 | 0.077 |
| RD expenditure | 7.010 | 7.097 | −1.849 | 0.065 | 4,496 | −0.180 | 0.005 | 0.087 |
| Assets | 9.613 | 9.659 | −0.876 | 0.381 | 4,529 | −0.150 | 0.057 | 0.046 |
| Employees | 3.805 | 3.763 | 0.871 | 0.384 | 4,450 | −0.052 | 0.134 | 0.041 |
| Cost of goods sold | 8.560 | 8.589 | −0.454 | 0.650 | 4,507 | −0.153 | 0.095 | 0.029 |

behavior. Table 10 performs t-tests for differences in means for these firm-level patenting characteristics. The results show small differences in the number of patents for each, but in opposite direction, with leaders having more issued patents and followers having more applications.

**Table 10: Firm-level patent characteristics. Covariates logged.**

| Variable | Leaders | Followers | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Prior apps. filed (2 yrs) | 6.678 | 6.752 | −1.357 | 0.175 | 4,555 | −0.181 | 0.033 | 0.074 |
| Prior apps. filed (5 yrs) | 7.385 | 7.430 | −0.775 | 0.438 | 4,559 | −0.159 | 0.069 | 0.045 |
| Prior pats. issued (2 yrs) | 5.733 | 5.588 | 1.796 | 0.073 | 4,565 | −0.013 | 0.303 | 0.145 |
| Prior pats. issued (5 yrs) | 6.358 | 6.156 | 2.318 | 0.020 | 4,564 | 0.031 | 0.373 | 0.202 |

Finally, Table 11 compares the characteristics of the leader and follower patents in the sample that is, the applications that result in winning or losing the patent race. Leader and follower patents exhibit no significant difference in the number of claims at filing. However, leaders do have slightly longer first claims on average. Because longer claims are associated with narrower (i.e. less good) patent scope (Kuhn and Thompson 2017), this difference suggests that the follower patents are broader. This difference likely reflects the higher abandonment rate for follower patents, where those follower patents with the smallest scope would be more likely to be abandoned, leaving a pool with slightly broader scope remaining in the sample.

Table 11 also shows that leader patents have slightly more inventors listed on their patents. The (unlogged) economic magnitude of this effect is 0.17 inventors. This difference could also be a result of the differential abandonment rates, or could indicate that firms that devote more inventive resources to a project are (slightly) more likely to win patent races. Figure 19 shows the distribution of inventors, and Figure 20 shows the dose-response curve for this effect.

**Table 11: Patent characteristics. Covariates logged.**

| Variable | Leaders | Followers | t | p | df | lower | upper | difference |
|---|---|---|---|---|---|---|---|---|
| Number of claims at filing | 3.147 | 3.159 | −0.623 | 0.533 | 3,490 | −0.051 | 0.027 | 0.012 |
| Length of first claim at filing (words) | 4.671 | 4.464 | 12.021 | 0 | 3,607 | 0.174 | 0.242 | 0.208 |
| Number of inventors | 0.874 | 0.816 | 2.828 | 0.005 | 3,401 | 0.018 | 0.098 | 0.058 |

Taken together, Table 8, Table 9, Table 10, and Table 11 suggest that firms that win patent races in the sample look very similar, both on overall firm observables and patent-

ing behavior, to those that lose patent races. The regression analysis controls explicitly for those small differences that remain in the 6-month sample. Insomuch as there remain any differences in samples induced by the differential abandonment rates, they will work against finding results for the key findings (e.g. follow-on research).

The finding that firms that win patent races are similar to those that lose them may seem counter-intuitive. In the broader world, firms with more resources, technical expertise, etc. would probably be more likely to win patent races. Many of such races, however, should be sufficiently lop-sided that the less able firm fails to file a comparable invention within 6 months of their abler competitor. In such instances, these races would not appear in the data. By construction, the regression discontinuity approach focuses on patent races that are closer, and thus where this balance would be expected.

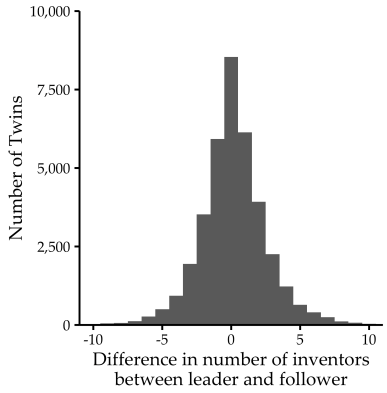



**Figure 19: Histogram showing the incidence of differences in the number of inventors between the leader and follower. The full sample of patented twins.**

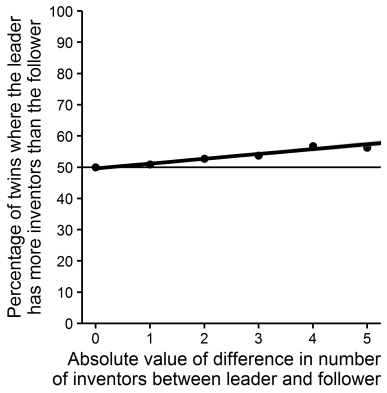



**Figure 20: Line plot showing the percentage of twins by the difference in inventors between the leader and follower. The full sample of patented twins.**

Because of the regression discontinuity approach, the estimates should be interpreted as local average treatment effects, centered on patent races where the competition between firms is credible enough that neither firm drops out before filing a patent application. Within this group, good balance is observed on observables, which provides support for the as-good-as-random assumption regarding the timing of patent filings.

### Firm Portfolio Approaches to Patent Racing

A key implication of having as-good-as-random allocation of patent races winners in the 6-month window is that the data should not show firms systematically winning or losing patent races. That is, if IBM were to win all its patent races in this window, that would argue strongly against the approach for getting causality. This prediction is tested explicitly on the top 100 patenters in the sample. If the as-good-as-random assumption is correct, the distribution of firm victories of patent races should reflect repeatedly drawing [win, lose] with 50:50 odds, a number of times equal to the number of races they participate in. That is, the modal firm should win 50% of their patent races and there should be a binomial

distribution around it. Figure 21 shows the empirical distribution of the victories of the top 100 firms in the solid line, and the corresponding binomial distribution, as simulated with 10,000 replications, in the dotted line.

The simulated curve replicates remarkably well the variation in firms' success at winning patent races. The only portion that it fails to replicate are the extreme tails of the distribution, suggesting that in the real world firms win or lose patent races less systematically than a binomial model would imply. This would likely reflect competitive pressures in the market where firms that are systematically losing invest more (or drop out), whereas those that are systematically winning have competitors that invest more (or drop out). These dynamics would produce the smaller dispersion visible in Figure 21.

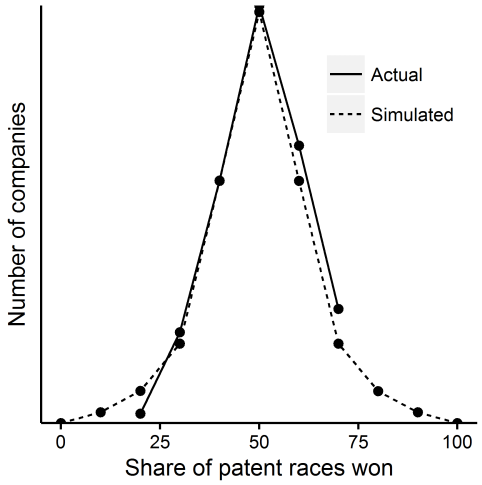



**Figure 21: Firm patent race winning percentage for the top 100 firms. The simulation curve plots the result of replicating 10,000 times a random draw [win, loss] with 50:50 odds of winning a race for each patent held by the top 100 firms.**

This test suggests that, conditional on a patent race being close enough that both firms file within 6 months of each other, most of the remaining difference is chance, providing support for the identifying assumption.

### Causal Effect of Winning a Patent Race

This section examines the causal effect of winning a patent race on firm behavior at three levels:

(1) the extent of patent protection actually obtained and maintained,

(2) the changes in firm-level outcomes (e.g. R&D), conditional on (i), and

(3) the follow-on research done by firms (as measured by patenting), conditional on (i).

Throughout Section 3.6, key results will be shown in two ways: (1) unconditional means, and (2) estimates adjusted for any remaining covariate imbalance via the following regression form (where the level of the analysis will depend on the estimand):

$$Outcome_i = \alpha_i + \beta_i Leader_i + \gamma_i Control_i + \epsilon_{ij} \quad (1)$$

Here the coefficient of interest is $\beta$, the effect of winning a patent race (i.e. being the "Leader"). Control variables include filing year fixed effects, Technology Center fixed effects, and controls for variables with imperfect balance (notably: the number of inventors). Standard errors are clustered at the firm x year level.

### Patent Protection Obtained and Maintained

By virtue of losing a patent race, a follower firm should be more likely to (1) abandon their patent (because it will have less value absent the contested claims), (2) get less scope in their final patent protection for ones that do issue, and (3) take longer in the prosecution of their patent. Conditional on issuance, follower firms should be less likely to keep their patents in force by paying the maintenance fees required in the third and seventh year after issuance (regressions 4 and 5). Figure 22 presents the unconditional means on these variables and Table 12 presents the regression results.

**Table 12: Patent protection and maintenance results**

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Abandonment (pp.) | Claim Narrowing (words) | Pendency (years) | Renewal Fee 1 (pp.) | Renewal Fee 2 (pp.) |
| | (1) | (2) | (3) | (4) | (5) |
| Is Leader | −0.133*** | −14.866*** | −0.693*** | 0.895 | 3.759* |
| | (0.023) | (3.556) | (0.049) | (0.823) | (1.943) |
| Sales at filing (logged) | | 0.653 | 0.178*** | 2.139* | −0.227 |
| | | (5.833) | (0.052) | (1.220) | (2.733) |
| R&D at filing (logged) | | 0.858 | 0.015 | 1.268 | 4.999** |
| | | (4.607) | (0.044) | (0.961) | (2.428) |
| Employees at filing (logged) | | 2.570 | −0.161*** | −5.787*** | −8.983*** |
| | | (6.134) | (0.054) | (1.375) | (2.799) |
| Earlier-filed patents (2 years, logged) | | −3.171 | −0.099*** | −1.196* | −1.226 |
| | | (2.560) | (0.029) | (0.633) | (1.394) |
| Number of inventors (logged) | | −0.147 | 0.147*** | −1.166 | −0.122 |
| | | (3.089) | (0.031) | (0.790) | (1.898) |
| Observations | 215,235 | 2,775 | 3,681 | 3,517 | 1,537 |
| $R^2$ | 0.290 | 0.435 | 0.929 | 0.917 | 0.787 |
| Adjusted $R^2$ | 0.290 | 0.429 | 0.928 | 0.917 | 0.783 |
| F Statistic | 3,386.712*** | 81.349*** | 1,642.735*** | 1,385.623*** | 232.610*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Model 1 run for all twins and includes filing year fixed effects.
Models 2-6 run for Compustat 6 and include technology center and filing year fixed effects.
Standard errors are clustered at the level of firm x filing year.

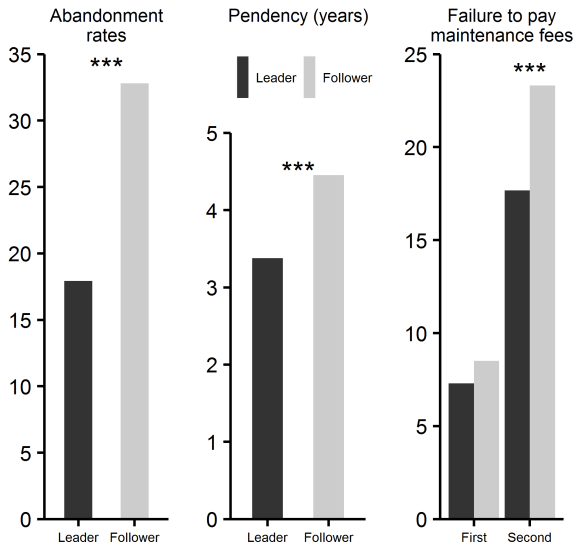



**Figure 22: Histogram showing the incidence of differences in the number of inventors between the leader and follower. The full sample of patented twins.**

Column 1 confirms that 13.3% fewer leader patents are abandoned than follower patents. For those patents do issue, column 2 shows that follower patents get their scope narrowed more during patent protection. Follower patents have, on average, 14.9 more words added to the first independent claim. As discussed by Kuhn and Thompson (2017), the addition of words to a claim typically narrows the scope of patent protection. Thus, not surprisingly, follower patents are narrowed more during examination.

Column 3 shows that the follower patents take 0.7 years longer to issue. This delay is presumably due to additional back and forth between the examiner and the follower firm's lawyers to try to revise the claims in light of the conflict, and likely implies both higher legal costs for the follower firm as well as less ability to enforce the patent, as the claims are not issued until later.

Finally, columns 4 and 5 show the renewal rates of patents at the first and second maintenance payments. There is no significant difference in renewal rates at the first maintenance period, with both leader and followers renewing about 92% of the time  perhaps because this is the lowest fee, at $1,600. By the second maintenance fee, this difference has grown, and leader patents are 3.8pp more likely to renew—although again against a backdrop of very high renewal rates. Indeed, it is telling that the renewal rate for all patents is 85.3% for the first maintenance fee and 64.4% for the second, so the rates are significantly higher in patent races.

Overall, the results show strong and consistent evidence that losing a patent race results in a less valuable potential patent and thus that applicants are more likely to abandon that patent, either during the prosecution process or later.

### Effect on Firm Level Outcomes

As already shown in Figure 14 and discussed in Section 5.4, larger firms win and lose patent races with roughly equal frequency. As such we should expect little-to-no-difference

in firm-level outcomes from winning a patent race. Table 13 confirms that this is indeed the case:

**Table 13: Firm-level financial results.**

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Sales | COGS | Pre-tax Income | R&D | Employees | Profitability |
| | (D.V.s 1-5 as an increase in logged variable) | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Is Leader | −0.006 | −0.004 | −0.036 | −0.004 | −0.007 | −0.002 |
| | (0.012) | (0.013) | (0.033) | (0.012) | (0.011) | (0.002) |
| Sales at filing (logged) | 0.097*** | 0.121*** | −0.047 | 0.129*** | 0.160*** | −0.015*** |
| | (0.031) | (0.033) | (0.100) | (0.031) | (0.028) | (0.006) |
| R&D at filing (logged) | 0.007 | 0.030 | 0.031 | −0.033 | −0.059*** | −0.007 |
| | (0.020) | (0.025) | (0.059) | (0.022) | (0.020) | (0.004) |
| Employees at filing (logged) | −0.158*** | −0.206*** | 0.005 | −0.154*** | −0.152*** | 0.018*** |
| | (0.028) | (0.034) | (0.076) | (0.028) | (0.024) | (0.006) |
| Earlier-filed patents (2 years, logged) | −0.005 | −0.016 | −0.015 | −0.018* | 0.013 | 0.005** |
| | (0.012) | (0.014) | (0.025) | (0.010) | (0.009) | (0.002) |
| Number of inventors (logged) | 0.008 | 0.006 | −0.001 | 0.004 | 0.006 | 0.003* |
| | (0.008) | (0.011) | (0.022) | (0.010) | (0.008) | (0.002) |
| Observations | 3,143 | 3,144 | 2,166 | 3,135 | 3,093 | 3,140 |
| $R^2$ | 0.594 | 0.530 | 0.437 | 0.460 | 0.391 | 0.093 |
| Adjusted $R^2$ | 0.590 | 0.526 | 0.430 | 0.455 | 0.385 | 0.085 |
| F Statistic | 156.887*** | 121.325*** | 57.252*** | 91.091*** | 67.723*** | 11.062*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Outcomes measure within-patent changes between the filing date of the leader and the later of the issue date of the leader and follower.
All models run for Compustat 6 and include technology center and filing year fixed effects.
Standard errors are clustered at the level of firm x filing year.

This precise null result is a large departure from theory, which predicts large effects on R&D from winning a patent race (see, e.g., Fudenberg, Gilbert, Stiglitz and Tirole (1983)), because those models typically assume a single project per firm. In reality, the analysis confirms that they take a portfolio approach. Thus, to investigate the effects predicted by theory models, the analysis is narrowed to project-level outcomes.

### Follow-on Research, Conditional on the Issuance of Both Patents

Like much of the literature, this thesis investigates project-level research using citations by later patents as a way to infer related future research. However, the citation results are also validated using a separate, textual similarity-based measure. This analysis also treats citations differently than many articles to mitigate some of the weaknesses inherent in citation-based measures.

The correct interpretation of patent citations is unclear, with many authors offering very different versions (see Kuhn and Younge (2016), for a more detailed discussion). Although follow-on citations can sometimes reflect follow-on research in an area, but that they are equally likely to reflect differences in the citation behavior in patents. That is, differences can reflect either a change in the propensity to do further work, or changes in the propensity to cite. In some cases, these are indistinguishable. In other cases, the institutions around patenting suggest one interpretation or the other.

Two different measures of citations and three different sources of citations are considered in the analysis. The measures are the total number of citations accruing to a patent and

the percentage of patents that accrue at least one citation. The latter measure is included as a robustness check, to ensure that highly-cited patents do not mask differential changes to lightly-cited ones. These citations are of three types, which are labeled "self cite," "twin cite," and "other cite"). All three sources reflect citations from later patents, but the analysis is divided into those by the same firm ("self cite"), by the other firm in the patent race ("twin cite"), and by other firms not in the patent race ("other cite"). Table 14 shows the estimates for these effects.

From columns (1) and (5) it is clear that leader patents get cited more, to a level that is both economically and statistically significant. Most of this effect derives from others firms citing the leader patent more than the follower patent, columns (2) and (6). Although it is possible that this effect reflects the larger patent scope first enumerated by the leader patent, it could also be true that it reflects examiners becoming more reliant on the leader patent for demarcating future claims. That is, examiners become familiar with the patent, and thus use it for rejections more frequently, leading to it being inserted into the citation record more often. However, due to the complex interplay between examiner and applicant citations in the form of applicant cross-citing, the analysis does not attempt separate these effects. (Kuhn, Younge and Marco 2016).

**Table 14: Patent-level citation results.**

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Total (Logged) | Other (Logged) | Self (Logged) | Twin (Logged) | Has Any | Has Other | Has Self | Has Twin |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Is Leader | 0.813*** | 0.490*** | 0.138*** | 0.703*** | 0.323*** | 0.234*** | 0.109*** | 0.760*** |
| | (0.029) | (0.029) | (0.021) | (0.016) | (0.014) | (0.016) | (0.015) | (0.013) |
| Sales at filing (logged) | 0.043 | 0.055* | 0.039 | 0.016 | −0.021 | −0.008 | 0.029 | 0.014 |
| | (0.034) | (0.029) | (0.032) | (0.019) | (0.015) | (0.017) | (0.018) | (0.013) |
| R&D at filing (logged) | −0.017 | −0.006 | −0.063*** | 0.010 | 0.021* | 0.014 | −0.032** | −0.005 |
| | (0.027) | (0.026) | (0.022) | (0.012) | (0.013) | (0.014) | (0.015) | (0.009) |
| Employees at filing (logged) | −0.075** | −0.069** | −0.061* | −0.026 | 0.011 | 0.006 | −0.036* | −0.008 |
| | (0.034) | (0.031) | (0.032) | (0.017) | (0.013) | (0.017) | (0.019) | (0.012) |
| Earlier-filed patents (2 years, logged) | 0.036** | −0.004 | 0.079*** | −0.003 | 0.008 | −0.004 | 0.047*** | 0.0001 |
| | (0.017) | (0.015) | (0.016) | (0.007) | (0.008) | (0.008) | (0.008) | (0.005) |
| Number of inventors (logged) | 0.064*** | 0.047** | 0.065*** | 0.003 | 0.006 | 0.013 | 0.040*** | −0.012 |
| | (0.024) | (0.023) | (0.018) | (0.012) | (0.011) | (0.013) | (0.013) | (0.009) |
| Observations | 3,670 | 3,670 | 3,670 | 3,670 | 3,670 | 3,670 | 3,670 | 3,670 |
| $R^2$ | 0.732 | 0.587 | 0.291 | 0.650 | 0.819 | 0.674 | 0.351 | 0.774 |
| Adjusted $R^2$ | 0.730 | 0.584 | 0.286 | 0.647 | 0.818 | 0.672 | 0.347 | 0.772 |
| F Statistic | 368.785*** | 191.924*** | 55.359*** | 250.649*** | 611.039*** | 278.961*** | 73.119*** | 462.371*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Citations counted 6 years after filing.
All regressions run for Compustat 6.
All regressions are conditional upon issuance.
All regressions include technology center and filing year fixed effects.
Standard errors are clustered at the level of firm x filing year.

Columns (3) and (7) show that the company that wins the patent race is more likely to cite its own leader patent in the future than is the other firm to cite its follower patent. Because

both firms are already clearly aware of their own patents, it is more straight-forward to interpret this difference as indicating that the leader firm pursues more follow-on innovation by virtue of winning the patent race. Under this interpretation, the effect of winning the patent race is large, leading firms to be 11% more likely to do follow-on research on their patent, and to do 14% more follow-on research in that area.

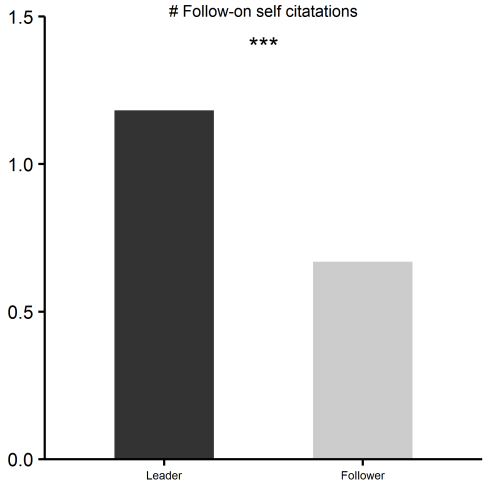



**Figure 23: Follow-on self-citations.**

Columns (4) and (8) focus on the citations by the competing firm in the patent race. They show that follower firms cite the leader patent 70% more than the leader firm cites the follower patent. This is unsurprising as the legal process of rejection makes the follower firm very familiar with the leader patent, but the converse it not true. Moreover, patent applicants are subject to a statutory duty of disclosure that requires them to acknowledge prior art they are familiar with. This duty frequently leads patent applicants to copy patent citations, such as references used to support a rejection, across groups of related patents. (Kuhn and Younge 2016). Thus, the increase in citations by the competing firm are most likely because of a change in citation behavior, rather than follow-on research behavior.

Having shown that firms that win a patent race do more follow-on research, the analysis also tests whether that follow-on research is more similar. This proposition is tested by directly applying a measure of the similarity between the patent twins and the follow-on work done by each firm. Firms winning patent races should have follow-on patents 'close' to the leader patent, whereas firms that lose the patent race will need to invent around, and thus will be 'further' from their follower patent (or the leader patent). Similarity is measured based on a cosine distance metric, calculated on the words in the description (aka specification) of each invention (see Younge and Kuhn (2015) for a detailed discussion of this metric). High similarity indicates that similar content words are used to describe both inventions, whereas low similarity indicates the usage of very different content words.

Figure 24 shows that, indeed, the content of the follow-on patenting by the leader is 'closer' to the patent that won the pace race than the follower's work is to their patent that lost the patent race. This relative decrease in proximity by followers suggests that many are abandoning that direct area of research and instead moving to adjacent areas. The right-most

part of this graph, showing very-high similarity, reflects patent divisionals or continuations. Not surprisingly, the leader firm files more divisionals / continuations, reflecting continuing interest in the specific area of the patent race win.

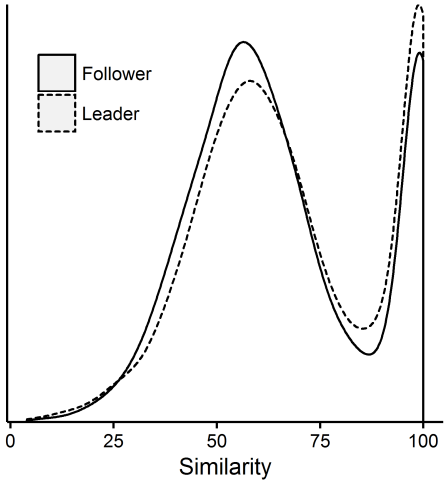



**Figure 24: Patent race leaders do follow-on research that is more similar to their winning patent, than do followers of their losing patents (Right-tail reflects divisionals / continuations).**

Thus, following a patent race victory, firms do more follow-on research. Further, the follow-on research that they do is more likely to be closely related to what was covered in their patent (and conversely that losers of patents races appear to be inventing around).

## Conclusion

This chapter provides a first large-scale look at how patent races happen in the real world. Using detailed administrative data from patent examinations and a novel approach for identifying patent races, it establishes key facts about who races and where, and provide causal estimates for the effects of winning a patent race.

We show that patent racing is common, with 10-11% of all patents in races, with large firms being more likely to race. In areas of technology that are usually thought of as having weaker patents (computers and communications) patent races are more common, whereas in areas with stronger patents (biotechnology and mechanical) they are less frequent. We also show that patents in patent races are much more likely than the average patent to have their scope narrowed during prosecution. Thus, patent races are more likely to occur in areas with a high-density of intellectual property protection per unit of "technology space". This could, for example, occur if firms in patent races were taking smaller innovative steps in order to be the first to file, and thus remained closer to the status quo, which one would expect to be more crowded with patents.

We also consider the causal effects of a patent race on those who win (the "leaders") and those who lose (the "followers"). To estimate these effects it uses a regression discontinuity approach, focusing on cases where both the leader and the follower file patents on the same invention in a 6-month window.

The results present strong evidence of firms that lose the patent race being more likely to abandon their application and more likely to let the protection lapse by not making maintenance payments. In support of theories that incorporate knowledge accumulation as a result of investing in a patent race, the results show that 67% of those that lose a race still receive patent coverage from their discovery.

Most importantly, we show that winning a patent race affects follow-on innovation, both in direction and magnitude. Contrary to many models that assume that firms participate in a single race, the results do not find this affect at the firm level and we show why one should not expect to observe such an effect because firms participate in a portfolio of races. Focusing down at the project level, however, the results demonstrate that leaders do significantly more follow-on research (14%) and that the work that they do is closer to the content of the original patent than is the follow-on research by followers.

Collectively, the results suggest that patent races are frequent and that they have important effects on the rate and direction of innovation by the firms. Moreover, the results provide empirical feedback on the assumptions and conclusions of patent racing models that can be used to model these better in the theory literature.

# Chapter 4

### Property Rights and Frictions in the Sale of Patents

## Introduction

The market for ideas, which includes the sale and license of patents, is an important part of how the patent system provides incentives for innovation. Moreover, ideas are a key input in a firm's production function, enabling the productive and competitive use of land, labor, and capital. Transactions for patent rights can thus spur economic growth and provide significant gains from trade by enabling the division of innovative labor, facilitating the exploitation of complementary assets, and avoiding the misallocation of resources. (Akcigit, Celik, & Greenwood, 2016; Arora, Fosfuri, & Gambardella, 2004; Serrano, 2011). However, a long line of research finds that the market for ideas suffers from substantial imperfections, and the market for buying and selling patents is thought to be particularly illiquid and inefficient. (Agrawal, Cockburn, & Zhang, 2015; Arora et al., 2004; Teece, 1986).

The prevalence of market frictions raises the concern that exchanges efficient from an institutional perspective may fail to occur, potentially reducing social welfare. On the supply side, market failures in technology transfers may prevent the exploitation of patented technology that a firm develops but does not commercialize. On the demand side, ownership fragmentation of complementary patents is associated with a reduced likelihood of in-licensing and more aggressive patenting within the firm. (Cockburn, MacGarvie, & Mueller, 2010; Ziedonis, 2004). Thompson and Kuhn (2016) found that over 10% of patents duplicate contemporaneous efforts by other firms, which may reflect in part the difficulty in transacting for technology. Even worse, the expectation of transactional difficulties may deleteriously affect market entry decisions and market structure. Indeed, imperfections in the market for ideas have been found to drive both firm commercialization strategies and inter-industry differences in firm boundary decisions. (Gans & Stern, 2003, 2010).

Patent scope is central to the sale of ideas. Awarding an inventor a patent on a new idea partially solves a commitment problem that would otherwise prevent the inventor from selling the idea. (Arrow, 1962). In the absence of a patent, a prospective buyer cannot credibly promise not to steal the idea should the inventor reveal it, while the inventor cannot credibly promise to reveal the idea should the prospective buyer pay for it. A firm's ability to use a particular patent to overcome this transactional hurdle derives from two factors: (1) the scope of the patent's legal right to exclude and (2) the effectiveness of that legal right in providing market exclusivity. In this chapter I employ a causal instrument to investigate the effect of an exogenous shock to the scope of a patent's legal right to exclude on the likelihood that the patent is sold. I then examine variation in the effectiveness of the legal right by interacting the instrument with endogenous firm, industry, and market characteristics.

This chapter contributes to a well-established literature on the transfer of technology via the market for ideas, the study of which is crucial for understanding the effects of the patent system on firm strategy and for crafting public policy to improve economic efficiency and provide incentives for innovation. Both patent sales and patent licenses can be used to transfer rights to technology, and a long line of research investigates the determinants of patent licensing. (See, e.g., Agrawal et al., 2015; Arora et al., 2004; Arora & Ceccagnoli, 2006;

Arora & Gambardella, 2010; Cockburn et al., 2010; Fosfuri et al., 2006; Mowery et al., 2001). However, because a patent license typically does not transfer to the recipient the right to sue those who infringe the patent, the determinants and consequences of sales and licenses often differ in various ways. Although the literature on the market for ideas provides important insight into transactions for patent rights, causal evidence is limited and the literature does not address the relationship between patent sales and a patent's legal right to exclude. This chapter investigates what is thus a fundamental and unanswered question—how, when, and under what conditions the scope of the right to exclude affects patent sales.

Patent sales have received less attention from scholars than patent licenses, but a bourgeoning literature highlights the importance of the secondary market for patents. Lamoreaux and Sokoloff (1999, 2001) showed that an active secondary market for patents in the late eighteenth and early nineteenth centuries gave rise to a large group of specialized independent inventors. Serrano (2010) presented descriptive evidence that patent sales are correlated with factors such as patent age, firm size, and forward citation counts. Hochberg, Serrano, and Ziedonis (2014) found a positive relationship between patent sales and the rate of lending to startups. Serrano (2011) quantified the gains to trade in the market for patents based on patent sales and renewal rates, while Akcigit et al. (2016) estimated a structural model and found that patent sales spur economic growth. Bessen, Meurer, and Ford (2011) argue that the explosion in patent litigation over previous decades is attributable in large part to non-practicing entities (NPEs) that purchase and monetize patents through licensing and litigation, but other research suggests that the patents acquired and asserted by NPEs are similar in quality and scope to other litigated patents. (Fischer & Henkel, 2012; Risch, 2011). Finally, Galasso, Schankerman, and Serrano (2013) show that while patent sales driven by comparative advantage in commercialization increase litigation, sales resulting from advantages in patent enforcement reduce it.

To investigate the role of property rights and frictions in patent sales, this chapter employs a range of novel measures and data sets developed by the author. The identification strategy is based on work by Kuhn and Thompson (2017), who develop a measure of a patent's legal right to exclude by analyzing the full text of the claims of more than 1 million patents. Because patent applications are quasi-randomly assigned to patent examiners within an area of technology and because patent examiners vary in their average effect on claim scope, a patent's examiner serves as a causal shock to the right to exclude provided by the patent relative to the underlying idea. To measure the effect of this shock, patent sales data were constructed by analyzing more than 4 million USPTO patent assignment agreements, which patent owners used to record changes to ownership rights. Measures of patent congestion, invention familiarity, and invention originality were constructed using patent-to-patent textual similarity data developed by Younge and Kuhn (2015, 2016), who employ a vector space similarity model to compare the full text of every patent to the full text of every other patent.

I show that receiving a broader patent increases the likelihood that the patent is sold, likely reflecting the fact that the sale of a broader exclusionary right provides greater gains to trade that are more likely to overcome the considerable transaction costs associated with patent sales. However, the magnitude of the effect depends substantially upon firm, technology, and market characteristics, likely reflecting variation in the practical effectiveness of patent rights across these dimensions for providing market exclusivity. In particular, the

scope of the right to exclude is more important for smaller firms and for patents in thicker and less congested markets. I also demonstrate that the between-industry variation in the effect of patent scope on patent sales aligns broadly with survey evidence described by Cohen, Nelson, and Walsh (2000) regarding industry-level patent importance and efficacy. Importantly, the likelihood that a patent is sold depends on the interaction of the realized scope of the legal right to exclude with the reported effectiveness of patents for exclusion in an industry. Likely for this reason, the right to exclude plays an outsized role in facilitating transfers in mechanical, chemical, and medical device industries and a relatively minor role in the semiconductor and electrical components industries.

This chapter makes several contributions to the literature on the market for ideas. First, it provides the first causal evidence of the relationship between patent sales and the scope of the legal right to exclude. Second, it demonstrates how the effectiveness of the right to exclude, and hence the strategic role patents play for firms, varies across firm and industry characteristics. Third, it links the market for patents with fundamental research on market design by illustrating interactions between market characteristics and the causal effect of the right to exclude. Fourth, it introduces new data and measures to provide current and comprehensive descriptive evidence of the extent and characteristics of the market for patents. Together these results have important positive implications for firm strategy since they illustrate ways in which a firm's ability to acquire, sell, and exploit innovation depend on market and intellectual property characteristics. At the same time, these findings have important normative implications for how the patent system may be modified to better incentivize innovation and allocate new ideas to where they may be most efficiently exploited.

Section 4.2 describes why understanding the determinants of patent sales is important for scholars of firm strategy and innovation economics. Section 4.3 develops testable hypotheses based on a discussion of the real-world implications of the right to exclude. Section 4.4 describes the empirical specification, including the econometric model, data and measures, and summary statistics. Section 4.4 also discusses the causal instrument used to investigate the effect of receiving a broader patent on the likelihood of a sale and introduces new measures of patent rights congestion, patent market thickness, technological originality, and technological familiarity. Section 4.5 presents the results from estimating the empirical model, which broadly support the hypotheses. Section 4.6 discusses the results and concludes.

## The Importance of Patent Ownership

Ronald Coase argued that a production input is incorporated into a firm when the transaction cost associated with coordinating production in the market, under imperfect information, is greater than within the firm. (Coase, 1937). In many cases, economies of scope mean that multiple products may be produced more efficiently via joint production using shared inputs. (Williamson, 1975). These forces are particularly salient in the development of new ideas, the production and commercialization of which frequently requires an array of complementary assets. (Teece, 1986). Joint production is optimally organized within the same firm if the economies of scope are based on the ongoing use of proprietary knowhow. (Teece, 1980, 1982). The importance of organizational knowledge explains why a firm needs access to many ideas, but not why the firm needs to own the ideas outright. Indeed, a belief that contracting is sufficient to facilitate the efficient flow of rights seems implicit in much of

the literature on the market for ideas, given that the literature to date has focused primarily on patent licensing. Nevertheless, several important distinctions between patent licensing and patent sales suggest that patent sales warrant independent investigation.

Most fundamentally, a license does not generally transfer the right to sue. Patent law is grounded primarily on negative, exclusionary rights rather than positive rights. For example, a patent does not provide a right to practice the claimed invention—a commercial product or process may be covered by potentially many different patents. Instead, a patent provides the right to exclude others from making, using, or selling the claimed invention. In the same way, a patent license transfers to the recipient not the right to sue, but rather the right to not be sued by the patent owner. However, patent law generally that any law suit alleging infringement of a patent be instigated by the patent's owner. A licensee only has standing to file a lawsuit without the patent owner's participation if the license is not only exclusive but also transfers "all substantive rights" associated with the patent to the license holder. Courts have interpreted these requirements broadly, holding that even field of use restrictions are enough to deny standing to the licensee. (*Luminara Worldwide, LLC. V. Liown Elecs. Co.*, Fed. Cir. Feb. 29, 2016). A licensee can in some circumstances initiate a lawsuit in conjunction with the patent's owner, but the difficulty of contractually resolving in advance the incentives and responsibilities inherent in a complex and prolonged patent infringement suit hinder such arrangements. For instance, the patent owner may have disincentives to incur costs, exert effort, sue its own actual or potential customers, or suffer negative publicity associated with a patent suit. Faced with a high and somewhat ambiguous bar to judicial standing, significant negotiation costs, and potentially conflicting incentives, a prospective licensee has significant motivation to instead acquire the patent outright if exercising the legal right to exclude is likely to be an important component of the licensee's strategy.

The United States Patent and Trademark Office ("USPTO") evaluates each patent application via a complex examination process, but rigorously examining each patent application to the point of certainty would be an inefficient use of patent office resources given that few patents are ever litigated. (Lemley, 2001). Patents are thus probabilistic in the sense that the validity, scope, and infringement of a patent are all unclear until each issue has been carefully litigated, and different actors may estimate these probabilities differently. (Lemley & Shapiro, 2005; Merges, 1999). The probabilistic nature of patents exacerbates the contracting costs of negotiating a license and underscores the importance of patent litigation. In the presence of uncertainty, rights holders may choose to establish a reputation for toughness by demonstrating the willingness and ability to engage in suit. (Agarwal, Ganco, & Ziedonis, 2009). A reputation for toughness can only be established on the basis of a meaningful threat.

In practice, very few patents are ever litigated, and scholars frequently study other uses of patents such as signaling, collaboration, and knowledge transfer. Nevertheless, the efficacy of these non-litigation uses for patents is grounded in the threat of litigation since patents encourage the free exchange of information by discouraging appropriation without consent. Acquiring the right to sue is an important motivator for patent sales because it facilitates a key strategy for many patent holders. Faced with a fragmented web of property rights spread across many different owners, a firm intending to commercialize in an area of technology may need to acquire patents for defensive purposes. The firm can then respond to litigation, actual or threatened, with a countersuit using its own patents. The threat of mutual holdup

facilitates a faster resolution to the potential standoff. (Somaya, 2002, 2003, 2012).

The differences in the legal effect of patent sales and patent licensing drives a distinction in their strategic implications. Both are ways for the original patent owner to generate revenue. However, patent sales are in theory driven by the most efficient owner of the right, while patent licensing is in theory driven by the most efficient use of the right. For instance, licensing may be an appropriate strategy when a patent protects a general purpose technology with applications in a variety of non-competing contexts that are not embodied in a single firm. In contrast, if a patent protects an idea whose exclusive use would provide a significant competitive advantage but whose development would require extensive investment and a range of complementary assets, then the most efficient course of action may be to sell the patent to the firm best equipped to exploit it. I do not observe licensing and therefore cannot directly investigate the trade-offs between licensing and assignments in this chapter. Nevertheless, as discussed in the Section 2.2, the extensive literature on patent licensing provides important insights that inform my analysis of patent sales.

### Institutions and Hypotheses

The patent system exists in part to solve a fundamental commitment problem that hinders the sale of new ideas. In the absence of patent protection, the prospective buyer cannot commit to paying for an idea if the idea is revealed before payment is received, while the prospective seller cannot commit to revealing the idea if payment is received before the idea is revealed. (Arrow, 1962). A patent reduces the danger of expropriation by granting the patent holder a tradable right to exclude others from using the idea. However, the patent system provides only an imperfect solution to the danger of expropriation, for two reasons. The first, addressed in Section 2.2.1, is that the scope of the legal right to exclude provided by the patent may fail to completely encompass the underlying idea. The second, addressed in Section 2.2.2, is that the practical effectiveness of the legal right to exclude varies with firm, industry, and market characteristics.

### The Delineation of the Right to Exclude

To secure a patent, an applicant files a patent application that includes both a detailed description of the invention and a set of claims that delineate the requested legal right to exclude. The USPTO examines each patent application both to weed out unwarranted applications and to ensure that patents are not granted with overly broad exclusionary rights. In practice the decision as to whether to grant a patent at all affects only patent applications on marginal ideas. After all, about 75% of patent applications ultimately issue as patents, and of the ones that do not issue nearly half are abandoned for non-technological reasons such as financial constraints. (Lemley & Sampat, 2008). In contrast, the delineation of the patent's right to exclude is the defining outcome of the patent examination process. (Kuhn & Thompson 2017). Indeed, for most patent applications the examination process takes the form of a negotiation between the patent applicant and patent examiner over the scope of the patent's right to exclude (i.e. the patent's claims).

In a typical patent examination, the patent examiner rejects the applicant's initial claims on the basis that the claimed idea is either not new or is obvious in view of existing inventions.

Although the detailed description of the idea is fixed at the time of filing the patent application, the patent applicant can usually overcome the initial rejection by amending the claims to encompass a narrower exclusionary right. This process of rejection and response may be repeated indefinitely until the examiner agrees to issue the patent or the applicant abandons the applications. In practice, about 70% of all patents are granted with claims that are narrower than those originally filed with the patent application. (Kuhn & Thompson 2017). Importantly for this paper's identification strategy, patent applications are quasi-randomly assigned to patent examiners within a particular area of technology, and patent examiners vary considerably in the degree to which they cause applicants to narrow their claims. (Kuhn & Thompson 2017).

The literature offers few clear theoretical predictions about the effects of receiving a broader or narrower right to exclude. Merges and Nelson (1994) argue that awarding broad patent claims may hinder technological progress, but that the effects will vary based on underlying features of the technology such as the extent to which rapid innovation requires diverse information inputs. However, they do not predict the effect of scope on the likelihood that a patent is sold. Empirically, Gans et al. (2008) show that resolving uncertainty about the scope of a patent under examination increases the likelihood that the patent will be licensed, but whether these licenses resulted from a reduction in information asymmetries or a higher-than-expected realized patent scope is unclear.[1] Further, sample size constraints and fundamental differences between licensing and sales limit the applicability of their result to the sale of patents across firms, time, and industries.

Prospective buyers and sellers must overcome significant transaction costs to achieve a patent sale. In addition to the transaction costs associated with incomplete and uncertain appropriability, patent transactions also involve non-trivial search, bargaining, and enforcement costs. A prospective buyer and seller must find each other, convey the idea, ascertain the characteristics of the patent that protects the idea, and negotiate the purchase of the patent rights. The dollar value of these costs vary widely across deals, firms, technologies and industries, but Teece (1977) estimated the total cost of an inter-firm technology transfer at between 2% and 59% of the value of the technology in an early study of 27 projects.

The key to understanding the effect of an increase in a patent's scope on the likelihood that the patent is sold lies in a return to the commitment problem the patent system is in part designed to solve. As part of the random process of discovery, a firm produces a variety of ideas. The firm may decide to commercialize some of these ideas but not others. Importantly, the lengthy time period required to secure patent protection and the difficulty in measuring the right to exclude mean that the decision to commercialize often must be made independently of realized USPTO outcomes. If the firm has decided to commercialize

[1]It should be noted that patent claims are subject to several types of uncertainty. Gans et al. (2008) investigate the effect of resolving one type—what the claims are—which is not determined until the patent office indicates to the applicant that patent application is ready to be issued. Of course, even after a patent is issued, uncertainty remains regarding both whether a court would invalidate the claims (e.g., if new prior art is unearthed) and whether the claims are being infringed. Applicants deal with this second type of uncertainty in part by including in the patent "dependent" claims that are narrower than the patent's main, independent claims and that serve as fallback positions in the event that the independent claims are invalidated. Because of these fallback positions, an applicant granted a broader patent generally does not give up the narrower protection, which means that a broader patent is strictly better for the patent owner than a narrower one even if broader claims are at greater risk of being invalidated. (Kuhn & Thompson 2017).

the idea, then presumably it will retain the patent rights regardless of the specific right to exclude afforded by the patent, although the firm would certainly prefer to receive a broader exclusionary right. If instead the firm has decided not to commercialize the idea, then the patent that protects the idea likely has limited value to the firm and may be a candidate for sale.

Of course, a firm can employ a patent offensively even in the absence of commercialization, but in practice most firms focus on competing in the product market rather than litigating patents on ideas they have chosen not to pursue. Firms that litigate must not only incur large up front costs but also risk retaliatory countersuits based on the defendant's own patents. Such costs and risks are likely unwarranted without the added benefit of product-market exclusivity that can only derive from actually practicing the patented invention. Some non-practicing firms exclusively pursue strategies focused on licensing and litigation, but these firms tend to purchase rather than sell patents.

If patent's original owner would like to sell it, then the likelihood of that sale occurring may depend in part upon how effectively the right to exclude afforded by the patent resolves the commitment problem. If the right to exclude is narrow, a prospective purchaser could commercialize the idea without actually purchasing the patent by tailoring the commercial product to avoid infringing the patent's narrow exclusionary right (i.e. "designing around" the patent). Alternately, the prospective purchaser of a narrow exclusionary right might worry that purchasing the patent would secure insufficient protection to prevent competitors from designing around the patent. In either case, gains to trade provided by the sale of a broader patent would be more likely to exceed the considerable transaction costs typically associated with patent sales. These observations lead to the first hypothesis.

> **Hypothesis 1.** *Holding fixed the underlying idea, a patent with a broader exclusionary right is more likely to be sold.*

### The Effectiveness of the Right to Exclude

The practical effectiveness of patents for different purposes such as market exclusion or cross-licensing varies widely between industries (Cohen et al., 2000), and patent licensing rates reflect this variation. For instance, licensing rates are correlated with industry-reported patent effectiveness. (Anand & Khanna, 2000; Arora & Ceccagnoli, 2006). Industry-level variation in patent effectiveness may in part reflect fundamental differences in technology. Industries differ in terms of product development lifecycles, technological opportunities to "design around" a given invention, the number of distinct ideas that tend to be embodied in a given product or process, the ways in which inventions are translated from ideas to language, and many other characteristics. At the same time, variation in patent effectiveness may also be driven by industry structure, which itself may be endogenously related to technological characteristics. Regardless of the cause of this differential effectiveness, the result is likely the same. In industries where patents are reported as being more effective for exclusion, the precise delineation of the right to exclude is likely more important than in other industries for providing market safety and thereby facilitating sales. These observations lead to the second hypothesis.

**Hypothesis 2.** *Holding fixed the underlying idea, the effect of the scope of a patent's exclusionary right on the likelihood that the patent is sold is greater in industries where patents are reported as being more effective for exclusion.*

The likelihood that a good is sold depends not only upon the characteristics of the good itself, but also on the market in which the prospective sale would occur. According to the market design literature, an efficient market must exhibit thickness, a lack of congestion, and market safety (Roth, 2008), all of which have been shown to correlate with an increased likelihood of patent licensing (Agrawal et al., 2015). Indeed, as discussed in Section 2.2.2, patent markets seem particularly at risk of market design problems. Market safety is the degree of risk incurred by participating in a market. The risk incurred by participating in the market for ideas is that of expropriation, the prevention of which is the very purpose behind the right to exclude provided by the patent system. In line with Hypothesis 1, the right to exclude thus serves to provide market safety, ensuring that the buyer actually receives the right to exclude purportedly provided by the patent.

Under a classical definition, market thickness is directly proportional to both the number of buyers and the number of sellers in the market. (Roth, 2008). However, the market for patents functions quite differently than conventional markets. (See, e.g., Niioka, 2006; Richardson, Oliver, & Costa, 2016). The supply of patent rights is relatively high, since the random nature of technological discovery means that many firms have produced patents that are of limited value to the firm and that they would be happy to sell. On the other hand, the demand for patent rights is naturally low, in part because only a limited number of firms are equipped to make use of patents in a particular area of technology and in part because of the imperfect exclusion provided by patent protection. Finally, transactions are often conducted through brokers since search and negotiation costs are often relatively high and since prospective buyers prefer anonymity to avoid strategic reactions by other firms. For these reasons, market thickness for patents depends not on the number of sellers but rather on the number of active buyers who have demonstrated a willingness to purchase patent rights.

A thin market may be related in both cause and consequence to the effect of the right to exclude on the likelihood that a patent is sold. In a thin market, the dearth of active buyers makes asset pricing unstable, and patents are difficult assets to price in the best of markets. Thus, in a thin market, variation in the right to exclude may have little impact on the patent's value to a prospective purchaser. At the same time, a lack of market thickness may indicate that patents are ineffective at exclusion regardless of how the right to exclude is delineated. In either case, the effect is the same and leads to the third hypothesis.

**Hypothesis 3.** *Holding fixed the underlying idea, the effect of the scope of a patent's exclusionary right on the likelihood that the patent is sold increases with market thickness.*

Market congestion arises from buyers' inability to evaluate (and hence to price) a surfeit of offers. When evaluating an offer requires a non-trivial amount of time, too many competing offers can destablize asset prices. (Roth, 2008). Under this classical definition, market congestion should not be problematic in the secondary market for patents since the market

does not feature offers to sell and is instead driven primarily by the demand side. However, in practice the market for patent rights suffers from congestion of a different sort.

As the number of patents issued has exploded, scholars have become increasingly concerned about patent thickets, which are generally defined as a web of overlapping patent rights in an area of technology. Patent thickets are associated with a decreased frequency of in-licensing (Ziedonis, 2004) but an increase in in-licensing expenditure (Cockburn et al., 2010). Congestion also speeds litigation settlement (Galasso & Schankerman, 2008) and may deter product market entry (Hall et al., 2015). In a finding that implicates both congestion and thickness, a firm's in-licensing rate exhibits an inverted U-shaped relationship with the number of potential suppliers. (Fosfuri, 2006).

In a market with congested rights, an individual patent's legal right to exclude is likely less effective for product market exclusion. For instance, a firm seeking to exclude a competitor from the product market by way of a lawsuit would likely face a retaliatory countersuit that relies on the competitor's own patent with a similar exclusionary right. Further, a mass of overlapping patent rights by different firms may raise the likelihood that in litigation a court would invalidate a broad patent on the grounds that it was improvidently granted by the USPTO. When the right to exclude is ineffective for excluding competitors from the product market, it provides less market safety for overcoming the commitment problem inherent to the sale of ideas. These observations lead to the fourth hypothesis.

> **Hypothesis 4.** *Holding fixed the underlying idea, the effect of the scope of a patent's exclusionary right on the likelihood that the patent is sold decreases with patent congestion.*

Aside from technology and market characteristics, the role of the right to exclude in protecting an idea is also likely to vary across firms. Larger firms with more established patent portfolios may be able to rely partially on repeated interactions and a network of related patents to protect ideas. For instance, larger firms may simply cross-license all patents necessary to produce a product rather than sharply negotiating usage rights for individual patents. In contrast, the precise contours of a single patent's legal right to exclude may be quite important for a smaller firm with fewer ideas and a shorter history of interactions with other firms. These observations lead to the fifth and final hypothesis.

> **Hypothesis 5.** *Holding fixed the underlying idea, the effect of the scope of a patent's exclusionary right on the likelihood that the patent is sold decreases with firm size.*

### Methodology

Section 4.4.1 describes the sample selection and variables. Section 4.4.2 introduces the identification strategy. Section 4.4.3 presents descriptive statistics on patent sales, and Section 4.4.4 describes the econometric specification.

**Sample Selection and Data**

Table 15 reports summary statistics for variables. The data are drawn from several sources. Data on patent assignments were constructed based on raw transfer data downloaded from the USPTO. The identification strategy, a plausibly exogenous shock to patent scope described in Section 4.4.2, was developed in related work based on a textual analysis of patent claims. (Kuhn & Thompson 2017). Patent bibliographic data was parsed from XML files provided by the USPTO as described by Thompson and Kuhn (2016). Measures of technological originality, technological familiarity, and patent rights congestion were constructed based on patent-to-patent similarity data described and validated by Younge and Kuhn (2015).

The unit of observation is the individual patent, and the sample includes 346,802 patents filed between January 1, 2004 and December 31, 2006. The sample includes all patents filed within the indicated time range with several exceptions. Most notably, the sample excludes biotechnology patents because the causal instrument is not valid in biotechnology. Additional details regarding sample selection are described in Appendix B.1.

I constructed assignment-related variables by analyzing 4,188,639 assignments from the USPTO patent assignment database. Each assignment agreement transfers ownership of a patent from the existing owner to a new owner. Although patent owners are not required by law to record assignments at the USPTO, they face strong incentives to do so because an earlier assignment can be legally voided by a subsequent assignment unless the earlier assignment is recorded. In the data, an assignment record identifies the patent number whose ownership is transferred, the names of the assignor (seller) and assignee (buyer), and the date the assignment was recorded. Instead, the parties to an assignment are identified via self-provided name and address information, not unique identifiers. Because this identification information is replete with typographical errors and alternative spellings, disambiguation was required in order to assign a unique firm identifier to each party to an assignment. The process for disambiguating the assignment parties is discussed in Appendix B.2.

Under U.S. law, every patent is initially owned in undivided and equal part by its inventors. In practice, inventors are typically required to assign the rights of their inventions to their employers under an employment agreement. I refer to the initial assignment transferring rights from the inventor or inventors to a firm as the original assignment. This chapter introduces a novel classification scheme in which each transfer after the original assignment is classified, based on the logic represented in Figure 25, as a sale, a spinoff, a firm exit, or a firm renaming by analyzing the full population of assignment data. Specifically, I first determine for each assignment whether the buyer ever previously received ownership of another patent. If not, I treat the focal assignment as the seller's entry into the patent market. I also determine for each assignment whether the buyer ever subsequently received ownership of another patent. If not, I treat the focal assignment as the buyer's exit from the patent market. These entry and exit variables allow me to classify each assignment. If an assignment is neither a seller exit nor a buyer entry, then I classify it as a sale. If an assignment is a seller exit but not a buyer entry, I classify it as a firm exit by the seller. If an assignment is a buyer entry but not a seller exit, I classify it as a spinoff that creates the buyer. If an assignment is both an exit by the seller and an entry by the buyer, I classify it as a firm renaming and exclude the assignment from subsequent analysis. Appendix B.2 provides more

detail regarding the construction and validation of these classification categories.

**Table 15: Variable definitions and summary statistics.**

| | Variable | Level | Definition | Mean | SD |
|---|---|---|---|---|---|
| 1 | Reassigned | Patent | Dummy = 1 if patent is transferred post-grant | 0.057 | 0.231 |
| 2 | Reassigned - any sale | Patent | Dummy = 1 if patent is transferred post-grant in a sale | 0.041 | 0.198 |
| 3 | Reassigned - any exit | Patent | Dummy = 1 if patent is transferred post-grant in a firm exit | 0.007 | 0.083 |
| 4 | Reassigned - any spinoff | Patent | Dummy = 1 if patent is transferred post-grant in a spinoff | 0.015 | 0.121 |
| 5 | Reassigned - small sale | Patent | Dummy = 1 if patent is transferred post-grant in a sale of 20 or fewer patents | 0.015 | 0.122 |
| 6 | Reassigned - small exit | Patent | Dummy = 1 if patent is transferred post-grant in a firm exit of 20 or fewer patents | 0.005 | 0.069 |
| 7 | Reassigned - small spinoff | Patent | Dummy = 1 if patent is transferred post-grant in a spinoff of 20 or fewer patents | 0.007 | 0.083 |
| 8 | Firm size | Patent | The log of the number of patents issued to the firm in the five years prior to the focal patent | 4.492 | 3.236 |
| 9 | Forward citations (total) | Patent | The number of forward citations to the focal patent | 8.233 | 17.524 |
| 10 | Forward citations (self) | Patent | The number of forward self-citations to the focal patent | 0.809 | 3.166 |
| 11 | Forward citations (102) | Patent | The number of forward blocking (novelty) citations to the focal patent | 0.045 | 0.238 |
| 12 | Forward citations (103) | Patent | The number of forward blocking (non-obviousness) citations to the focal patent | 0.082 | 0.330 |
| 13 | Familiarity | Patent | The normalized maximum textual similarity between the focal patent and any prior patent by the same firm | 0.000 | 0.978 |
| 14 | Originality | Patent | The normalized maximum textual similarity between the focal patent and any prior patent by a different firm, times -1. | 0.000 | 0.999 |
| 15 | Market congestion | Art unit | The normalized mean maximum textual similarity between the focal patent and any prior patent application | 0.000 | 1.000 |
| 16 | Market thickness | Art unit | The logged ratio of unique buyers to patents | 0.000 | 1.000 |

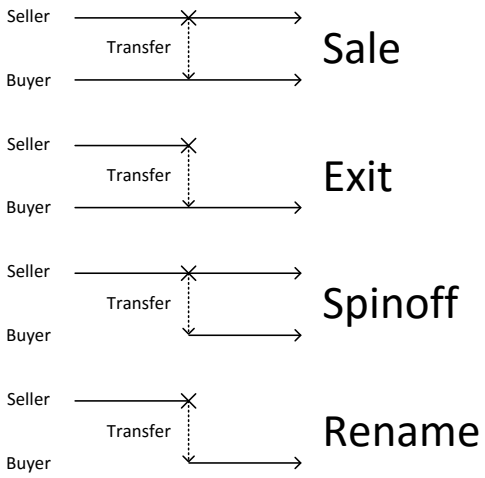



**Figure 25: Representation of the logic used to identify patent transfer type.**

The first variable in Table 15 indicates whether a patent was reassigned to another firm after it was issued to an initial firm. The second through fourth variables indicate whether the patent was reassigned post-grant in particular types of assignments, including sales, exits, and spinoffs. The fifth through seventh variables are similar but only count those transfers that include 20 or fewer patents between the buyer and seller in a given year. The cutoff of 20 is chosen arbitrarily, but conservatively, to focus attention on potentially strategic patent sales rather than the wholesale transfer of entire portfolios.

The data on assignments are analyzed subject to an important caveat. Multinational corporations sometimes engage in intra-firm patent transfers such as between domestic headquarters and foreign subsidiaries, often for organizational or tax reasons. (Arora, Belenzon & Rios, 2011). The current version of the data set does not identify such transfers. However, many such transfers occur before the patent is granted and as such do not affect the analysis of this paper. Further, transfers between firm subsidiaries are not arms-length market transactions and therefore should bias the estimates in the results toward zero.

The thirteenth and fourteenth variables in Table 15 were constructed based on patent-to-patent textual similarity data. Familiarity measures the extent to which the focal patent is similar to past work by the same firm. Originality measures the extent to which the focal patent is different than past work by other firms. Appendix B.2 provides further details regarding the construction of the familiarity and originality measures.

The fifteenth and sixteenth variables in Table 15 measure market thickness and rights congestion—characteristics of the local technology market in which a patent is situated. I define the local technology market based on the USPTO examining division (i.e. art unit) that examined the patent. Patent applications at the USPTO are assigned for examination to one of 531 art units based on the technology described and claimed in the application. The art units, which are split among eight Technology Centers, each includes a number of patent examiners skilled in examining a particular type of technology, such as mechanical presses (art unit 3725) or furnaces (art unit 3743).

Conventionally, market thickness is defined based on the number of buyers and sellers

making bids and offers. However, since the market for patents is driven primarily by demand rather than supply, as discussed in Section 4.4.1, I instead defined market thickness as the ratio of buyers to patents within the local technology market (i.e. art unit) of the focal patent. The raw value was logged and normalized to produce a z-score that provides a sense of the number of buyers to which a potential seller of a focal patent would have access. The interpretation of the thickness variable is that a patent having a thickness level of 1 is in an art unit that is 1 standard deviation thicker than the mean for all patents.

Conventionally, market congestion is a feature of markets in which offers must be evaluated in serial rather than in parallel and in which evaluating offers requires non-trivial time. However, because the market for patents features a surplus of rights rather than a surplus of offers, I instead measure the related concept of patent thickets. Scholars have proposed several measures of patent thickets, but all are based on patent citations or patent office technological classification. (Cockburn et al., 2010; Von Graevenitz, Wagner, & Harhoff, 2011; Ziedonis, 2004). In order to avoid many of the concerns about bias and selection inherent in patent citations and manual technological classification, I instead constructed a measure based on patent-to-patent textual similarity data. Specifically, the patent congestion variable is constructed by comparing the mean originality of all patents in the local technology market (i.e. art unit) to the originality of other patents in the sample. The interpretation of the congestion variable is that a patent having a congestion level of 1 is in an art unit that is 1 standard deviation more congested than the mean for all patents. Appendix B.2 provides further details regarding the construction of the congestion measure.

Table 16 reports summary statistics by broad industry category. As discussed earlier in Section 4.4, each patent is examined by an art unit at the patent office. The art units are divided into eight broad technology centers. In addition, I list as a separate category those patents issued by art units responsible for medical devices. As is shown in Table 16, the frequency of sales varies considerably by industry, with sales being most common in computer, networking, and electrical technologies and least common in chemical and mechanical technologies. The medical device, chemical, and computer industries are all thick and uncongested, suggesting that the right to exclude may be particularly effective in these broad categories. For the other industry categories, the mean congestion and thickness values are more ambiguous. For instance, mechanical category is moderately thick but also quite congested, while the electrical components industry is relatively uncongested but also quite thin.

**Table 16: Summary statistics by industry**

| Industry | Observations | Sale | Spinoff | Exit | Congestion | Thickness |
|---|---|---|---|---|---|---|
| All | 339,769 | 4.09% | 1.49% | 0.69% | 0.00 | -0.00 |
| Electrical | 53,368 | 5.45% | 2.09% | 0.69% | -0.18 | -0.43 |
| Semiconductors | 123,404 | 3.49% | 1.47% | 0.53% | 0.16 | -0.06 |
| Chemical | 43,990 | 2.51% | 1.30% | 1.03% | -0.61 | 0.12 |
| Computers | 30,199 | 11.30% | 1.26% | 0.79% | -0.58 | 0.15 |
| Miscellaneous | 39,745 | 2.27% | 1.37% | 0.48% | 0.40 | 0.26 |
| Mechanical | 41,605 | 2.27% | 1.13% | 0.49% | 0.52 | 0.12 |
| Medical devices | 7,458 | 4.16% | 2.15% | 3.35% | -0.45 | 0.72 |

### Identification Strategy

This paper employs as an identification strategy an instrumental variable that provides a plausibly exogenous shock to the final scope of the patent relative to the underlying idea. Patent applications are quasi-randomly assigned to patent examiners within an art unit, and patent examiners vary in the stringency with which they examine patent applications. (Cockburn, Kortum, and Stern (2002); Kuhn & Thompson (2017); Lemley and Sampat (2012); Sampat and Williams (2015)). The assignment of patent applications to patent examiners thus provides a natural experiment—patents examined by less stringent examiners provide on average a broader exclusive right relative to the underlying idea than those examined by more stringent examiners. Kuhn & Thompson (2017) show that the number of words in a patent's first independent claim is a good proxy for the patent's scope, and that the mean number of words added to the first independent claim in an examiner's patents (excluding the focal patent) is a good measure of examiner stringency.

Table 17 presents the first stage for all patents and for industry sub-samples. Were all examiners identical, a patent's scope would be determined entirely by factors such as the underlying technology and would not depend on examiner characteristics. However, Table 17 reveals considerable variation between examiners. A one standard deviation increase in examiner score predicts a .18 ($p<.01$) standard deviation increase in patent scope for all patents. Although the strength of the instrument varies somewhat by industry, examiner score remains a positive and highly significant predictor of patent scope.

**Table 17: First stage by industry.**

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Patent scope | | | | | | |
| | All | Electrical | Semicond. | Computers | Chemical | Mechanical | Medical |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Examiner score | 0.183*** | 0.274*** | 0.148*** | 0.246*** | 0.135*** | 0.162*** | 0.152*** |
| | (0.004) | (0.007) | (0.004) | (0.014) | (0.007) | (0.007) | (0.017) |
| Observations | 339,769 | 53,368 | 123,404 | 30,199 | 43,990 | 41,605 | 7,458 |
| $R^2$ | 0.046 | 0.085 | 0.029 | 0.089 | 0.022 | 0.043 | 0.042 |
| Adjusted $R^2$ | 0.043 | 0.083 | 0.028 | 0.083 | 0.017 | 0.040 | 0.036 |
| Residual Std. Error | 0.961 | 0.941 | 0.933 | 1.007 | 1.118 | 0.890 | 0.952 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All variables normalized as z-scores. All models run with application year by art unit fixed effects. Standard errors are clustered at the firm level.

A natural concern with any instrument is that although institutional factors suggest quasi-random assignment, the practical realities are such that assignment is not in fact random. Balance statistics indicate this is not the case. Kuhn & Thompson (2017) show that initial applications examined by lenient and stringent examiners are nearly identical on characteristics observable at the time of filing, and that patent examiner stringency varies little over time and by experience. Table 18 presents balance statistics comparing the applications examined by the upper and lower quartiles of examiners by stringency for the specific sample

analyzed in this chapter. The statistics show a slight selection effect due to fact that more stringent examiners are more likely to cause applicants to abandon low-quality applications. All of these differences are statistically significant due to the large sample size, but the differences are not economically significant. The means differ by at most .033 standard deviations, well below the threshold of .25 that would indicate a substantial difference. (See, e.g., Imbens & Wooldridge 2009). Further, because higher-quality applications are more likely to be transferred and because the small selection effect removes more low-quality applications from the set of applications examined by stringent examiners, the selection effect should slightly bias the results toward zero. Indeed, in unreported results, controlling for the variables included in Table 18 slightly increase the point estimates relative to the uncontrolled results.

**Table 18: Balance statistics (t-tests).**

| Variable | Broad | Narrow | T-Stat | Normalized Difference |
|---|---|---|---|---|
| Number of inventors | 2.544 | 2.554 | −1.259 | 0.004 |
| Number of words in patent | 7,323.451 | 7,669.633 | −9.707 | 0.033 |
| Number of claims | 20.717 | 20.794 | −0.747 | 0.003 |
| Number of words in first claim | 132.472 | 135.838 | −8.240 | 0.028 |

Observations: 169,880

Another concern with any instrument is violation of the exclusion restriction. Although patent examiners appear to be quasi-randomly assigned, a more stringent patent examiner may affect patents in ways other than the scope of the exclusive right. The highly-regulated, arms-length, and quality-controlled nature of the procedural rules governing patent examination should allay many of these concerns. However, tougher examiners may cause patents to issue more slowly, which could deter subsequent transfers. (Gans et al., 2008). In practice, the difference is relatively small—patents examined by a patent examiner one standard deviation tougher than the mean are delayed on average by about 2 months (Kuhn & Thompson, 2017)—but this concern cannot be ruled out completely. Another possibility, which again cannot be ruled out, is that examiner stringency may cause patent owners to perform increased or decreased follow-on innovation, which may affect whether or not the patent is sold. In part for this reason, the results include reduced form versions of all regression tables.

### Descriptive Statistics

Conventional wisdom suggests that patents are sold almost exclusively in large portfolios, which may render irrelevant the characteristics of any one patent. However, the data indicates that this is not the case. Figure 26 shows a patent-level histogram of first post-grant transfers, binned by the number of other patents included in the transfer. Although many transfers include 50 or more patents, the majority of patent transfers include 25 or fewer patents. Figure 27 plots the average examiner score of patents transferred post-grant as a function of the number of other patents included in the transfer. As would be expected, scope appears to plays a larger role in the transfer of small numbers of patents than in transactions for large portfolios. Figure 28 shows the frequency of different types of post-grant transfers by patent.

Clearly, most patents remain with the firm that produces them. However, about 5.7% of patents in the sample are transferred post-grant. Many of these transfers occur when large portfolios change hands, but about 1.5% of patents are transferred in small batches of 20 or fewer patents. To identify the role of patent scope as a driver of patent transfers, the results in the rest of the paper focus on transfers of patents that occur in groups of 20 or fewer. Although some patents are transferred more than once, such patents are quite rare, and subsequent analysis in this paper focuses on the first post-grant transfer for analytic simplicity.

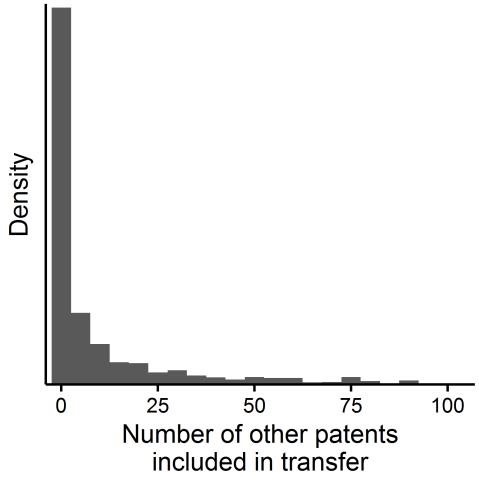



**Figure 26: The density of patent transfers by the number of patents included in the transfer.**

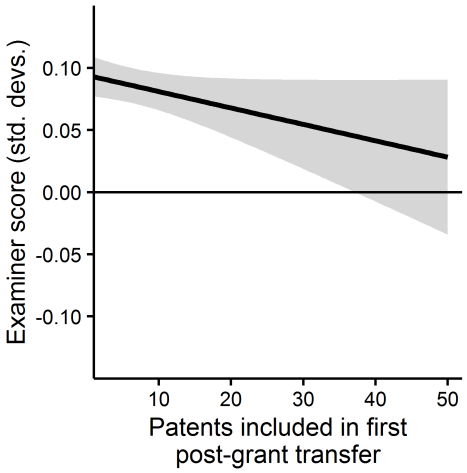



**Figure 27: Linear relationship between examiner score and the number of patents in a transfer.**

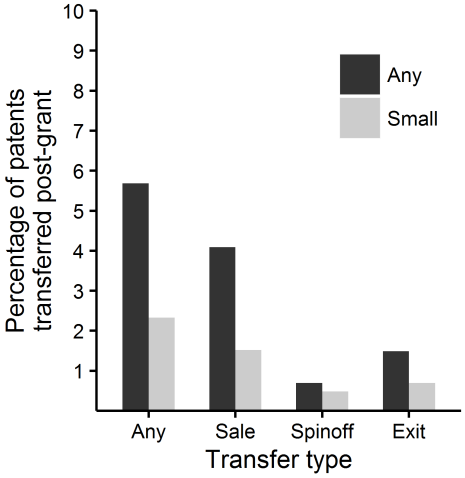



**Figure 28: Percentage of patents transferred. Small transfers include 25 or fewer patents.**

Most patent transfers are likely motivated by idiosyncratic characteristics of the underlying technology that are difficult to measure. Nevertheless, the data reveals patterns that

may shed light on the motivations behind some transfers. Table 19 presents the results of four models estimating the relationship between patent sales and patent citations made to the focal patent. Consistent with the findings of Serrano (2010), columns 1 and 2 show that patents that are sold receive 31% (p<.01) more total forward citations than patents that are not sold, but they receive 14% (p<.01) fewer citations by the firm that originally owned the patent. Thus, patents that are sold have on average a greater impact than those that are not, but unsurprisingly are less important to their original owners. Columns 3 and 4 show that patents that are sold are much more likely to be cited by patent examiners to reject future patent applications. Indeed, a patent that is sold is 2.3 (p<.01) times more likely to be cited in a novelty rejection than a patent that is not sold. Thus, patents that are sold are on average not only more impactful in a general sense, but also more strategically important in blocking future patent rights by other firms.

**Table 19: Sales and forward citations.**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Total (logged) | Self (logged) | Blocks others (102) | Blocks others (103) |
| | OLS | | Logistic (odds ratio) | |
| | (1) | (2) | (3) | (4) |
| Is sold | 0.305*** | −0.137*** | 2.312*** | 1.206*** |
| | (0.015) | (0.008) | (0.515) | (0.171) |
| Observations | 339,769 | 339,769 | 339,769 | 339,769 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All models run with application year fixed effects.

Another way to understand which patents are more likely to be transferred is to compare a focal patent's similarity to both prior work by the same firm and prior work by other firms. Table 20 presents the results of four models using a logistic regression to estimate the relationship between a patent's technological positioning using patent-to-patent textual similarity data and the odds ratio that a patent is transferred. The models split the sample by firm size into large firms and small firms.

A large firm seeking to establish and maintain a competitive advantage should retain the patents that contribute to such an advantage and sell the patents that do not. One way to establish a competitive advantage is to develop a technological competence that distinguishes the firm from its competitors. Accordingly, a large firm should be less likely to sell patents that are more central to the firm's area of competence and that are more economically original, since those are precisely the exclusionary rights likely to provide the greatest advantage. Columns 1 and 2 in Table 20 support such a story, suggesting that large firms are less likely to sell patents on technology familiar to the firm and less likely to sell patents on technology more original to the world outside the firm. Specifically, a large-firm patent one standard deviation more familiar to the firm is .84 (p<.01) times as likely to be sold, while a large-firm patent one standard deviation more original to the world is .89 (p<.01) times as likely to be sold.

**Table 20: Sales and technological relatedness.**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Odds ratio of sale | | | |
| | Large firms | | Small firms | |
| | (1) | (2) | (3) | (4) |
| Familiarity | 0.840*** | | 0.954* | |
| | (0.027) | | (0.025) | |
| Originality | | 0.888*** | | 1.136*** |
| | | (0.029) | | (0.022) |
| Observations | 123,807 | 123,963 | 80,462 | 110,609 |
| Log Likelihood | -5,610.591 | -5,634.137 | -8,254.129 | -12,675.660 |
| Akaike Inf. Crit. | 11,227.180 | 11,274.270 | 16,514.260 | 25,357.320 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All models run with application year fixed effects.

Patent sales by small firms are likely to be driven by different considerations. As noted in columns 1 and 2, patents on technology original to the world appear more valuable to large firms than patents on less original technology and thus in greater demand. For a small firm that seeks to monetize its patent portfolio but has relatively few patents to sell, the likelihood of a sale may be likely to be driven by demand-side considerations of precisely which ideas are valuable on the market. However, the extent to which a technology is unfamiliar to the firm may be less likely to influence whether a patent is sold for a small firm since small firms may be more willing to expand into unfamiliar technologies. Columns 3 and 4 in Table 20 support such a story, with small firms being much more likely to sell patents that are more original and with technological familiarity playing a minor role. Specifically, a small-firm patent one standard deviation more familiar to the firm is .95 (p<.1) times as likely to be sold, while a small-firm patent one standard deviation more original to the world is 1.136 (p<.01) times as likely to be sold.

**Econometric Specification**

Equation 1 provides the primary econometric specification. The likelihood that a patent is sold at any time after it is issued is estimated using a two-stage least squares regression. Each observation $i$ represents a single patent, while firms are indexed with $j$. In the first stage, a patent's scope is predicted based on the score of the examiner responsible for examining that patent. In the second stage, the estimated scope is used to predict the likelihood that the patent is sold. Dummy variables are included for each art unit during each patent application filing year to capture technology-level and year-level variation in the rate of sales.

$$\begin{aligned} Scope_i &= \eta * Score_i + \rho * ArtUnit_i * AppYear_i + \mu_i \\ Sale_i &= \beta * \widehat{Scope}_i + \gamma * \widehat{Scope}_i * C_i + \delta * C_i + \tau * ArtUnit_i * AppYear_i + \epsilon_{ij} \end{aligned} \tag{2}$$

The main specification does not include any additional control variables, but other specifications include a control variable $C_i$ that controls for firm size, market thickness, rights congestion, or patent effectiveness. In some secondary specifications, the dependent variable is replaced with a dummy indicating not whether the patent was sold, but rather whether the patent was transferred in a firm exit or spinoff transaction.

Patent scope (i.e. $Scope_i$) measures the breadth of the focal patent's right to exclude relative to other patents in the same area of technology. Similarly, $ExaminerScore_i$ measures the extent to which the focal patent's examiner tends to issue broad patent rights relative to the examiner's peers. Following Kuhn & Thompson (2017), $Scope_i$ and $ExaminerScore_i$ are defined on the level of the patent as described in Equations 2 and 3. The variable $Words$ measures the number of words in the first claim of the focal patent ($Words_i$), other patents in the art unit of the focal patent ($Words_a$), or other patents examined by the examiner of the focal patent ($Words_{ei}$). Importantly, Equation 3 is calculated for each patent with the focal patent excluded from the calculation to avoid introducing endogeneity into the instrument.

$$Scope_i = \frac{Words_i - mean\left(Words_a\right)}{StdDev\left(Words_a\right)} \tag{3}$$

$$ExaminerScore_i = \frac{mean\left(WordsAdded_a\right) - mean\left(WordsAdded_{ei}\right)}{StdDev\left(WordsAdded_a\right)} \tag{4}$$

Autocorrelation of the error term does not seem to be problematic since in unreported results the Huber-White robust standard errors are quite close to classical standard errors across all models. However, the error term in the regression could be correlated within any of various groupings. For example, firms could vary in their sensitivity to changes in patent scope. To address this concern, standard errors in ordinary least squares and two-stage least squares regressions were clustered at the level of the firm that originally owned the focal patent. Errors could be also correlated within patents examined by the same examiner or within patents examined by the same art unit. In unreported results, clustering at the level of art unit or patent examiner yields standard errors similar to clustering at the level of the firm.

As discussed in Section 4.4.2, violation of the exclusion restriction cannot completely be ruled out. For this reason, and because the dependent variable is binary with a relatively low mean (between 1% and 2%), I also estimated each model using a conditional logistic regression model in which the likelihood of a sale is regressed directly on examiner score. Equation 4 presents the specification for the conditional logistic regression. Estimating a logistic regression with fixed effects (i.e. a conditional logistic regression) can render estimates less consistent due to the incidental parameters problem. However, the number of fixed effects included Equation 3 is very small in comparison with the number of observations, and the number of observations within each group is quite large. Together these factors should minimize any bias introduced by including the incidental parameters in the model.

$$\log\left(\frac{Sale_i}{1-Sale_i}\right) = \beta * Score_i + \gamma * Score_i * C_i + \tau * TechCenter_i * AppYear_i + \mu_i \quad (5)$$

## Results

### Main Results

Table 21 presents estimates for naive models in which the likelihood that a patent is sold is regressed directly on patent scope. Model 1 presents the results for all patents, while models 2 through 21 present the results for industry subsamples. Broader patents are more likely to be sold, with a one standard deviation increase in patent scope associated with a .085 percentage point increase (p<.05) in the likelihood of a sale across the entire sample. The effect is positive but insignificant for all industries and is larger within the computer, chemical, and mechanical industries. However, patent scope is an endogenous result of the underlying invention, so the effects shown in Table 21 may be driven by characteristics of the underlying idea or the patent applicant rather than the scope of the exclusionary right.

**Table 21: Main results by industry, endogenous regression.**

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Likelihood of sale (percentage points) | | | | | | |
| | All | Electrical | Semicond. | Computers | Chemical | Mechanical | Medical |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Patent scope | 0.085** | 0.093 | 0.029 | 0.142* | 0.132 | 0.151 | 0.080 |
| | (0.042) | (0.077) | (0.088) | (0.079) | (0.082) | (0.103) | (0.175) |
| Observations | 339,769 | 53,368 | 123,404 | 30,199 | 43,990 | 41,605 | 7,458 |
| $R^2$ | 0.076 | 0.013 | 0.012 | 0.016 | 0.220 | 0.080 | 0.334 |
| Adjusted $R^2$ | 0.073 | 0.011 | 0.010 | 0.010 | 0.216 | 0.077 | 0.329 |
| Residual Std. Error | 11.445 | 10.578 | 11.598 | 9.672 | 12.115 | 11.984 | 15.050 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Covariate normalized as a z-score. All models run with application year by art unit fixed effects. Standard errors are clustered at the firm level.

To address endogeneity concerns, Table 22 presents estimates for reduced form ordinary least squares fixed effects models in which the likelihood that a patent is sold is regressed on the examiner score instrument. Examiner score has a larger effect than patent scope in Table 21, with a one standard deviation increase in examiner score increasing the likelihood that the patent is sold by .17 percentage points (p<.01). The largest effect is observed in the medical device industry, where a one standard deviation increase in examiner score increases the likelihood of a sale by 1.4 percentage points (p<.01). Moderate and statistically significant increases are also observed in the computer, chemical, and mechanical industries, while patent scope seems to exert little to no effect on sales in the electrical and semiconductor industries.

Table 23 presents the main results, a set of two-stage least squares regressions of the likelihood that a patent is sold on patent scope, with patent scope instrumented by the

**Table 22: Main results by industry, reduced form.**

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Likelihood of sale (percentage points) | | | | | | |
| | All | Electrical | Semicond. | Computers | Chemical | Mechanical | Medical |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Examiner score | 0.174*** | 0.074 | 0.040 | 0.275** | 0.291* | 0.405*** | 1.384*** |
| | (0.047) | (0.056) | (0.094) | (0.115) | (0.150) | (0.124) | (0.469) |
| Observations | 339,769 | 53,368 | 123,404 | 30,199 | 43,990 | 41,605 | 7,458 |
| $R^2$ | 0.076 | 0.013 | 0.012 | 0.017 | 0.220 | 0.081 | 0.338 |
| Adjusted $R^2$ | 0.073 | 0.011 | 0.010 | 0.010 | 0.217 | 0.078 | 0.333 |
| Residual Std. Error | 11.444 | 10.578 | 11.598 | 9.669 | 12.113 | 11.979 | 15.004 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Covariate normalized as a z-score. All models run with application year by art unit fixed effects. Standard errors are clustered at the firm level.

examiner score. The main effect is shown in Model 1, where a one standard deviation increase in patent scope increases the likelihood that a patent is sold by .95 percentage point (p<.01), a result which directly supports the main hypothesis.

The largest effect is observed in the medical device industry, where a one standard deviation increase in patent scope increases the likelihood of a sale by 9.1 percentage points (p<.01). Large and statistically significant effects are also observed in the computer, chemical, and mechanical industries. As noted in Table 16, the patent markets in chemical, computer, and medical device technologies are all thick and uncongested. While the mechanical sector is highly congested, it is also rather thick, which suggests a mature market in which buyers know how to value patents.

In contrast, patent scope has a negligible effect in the two thinnest industries: electrical component and semiconductor industries. Important semiconductor products are typically covered by many hundreds, and sometimes thousands, of different patents. Because of this fragmentation, the importance of the scope of any particular patent is likely to be negligible. The electrical component industry, including technologies such as television and telephony, is well-established and highly consolidated. In a market where the large majority of patents are owned by very large companies and are never transferred, the scope of a patent is unlikely to have much effect on the likelihood of sale.

The estimates presented in Table 22 are much larger and more significant than those presented in the endogenous regressions in Table 21, likely due to error in measuring patent scope. In the endogenous regression, the likelihood of a sale is regressed on the normalized number of words in its first independent claim, which is a statistically significant but very noisy measure of patent scope (Kuhn & Thompson, 2017). The examiner score provides a much less noisy measure by aggregating the patent examiner's impact on patent scope across all other patents issued by the examiner.

When interpreting the estimates in Table 22 and other instrumental variable regression results in this Thesis, the reader should keep in mind the magnitude of the effects in the first stage. Because a one standard deviation increase in examiner score increases a patent's

scope by about .2 standard deviations, a variation of one standard deviation in patent scope as indicated in the estimates in Table 22 is larger than most between-examiner variation observed in practice. For this reason, reduced form estimates regressing directly on examiner score are somewhat more interpretable than the instrumental variable results, so Table 25, Table 26, Table 27, and Table 28 present estimates of reduced form logistic regression models.

**Table 23: Main results by industry, instrumental variable.**

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Likelihood of sale (percentage points) | | | | | | |
| | All | Electrical | Semicond. | Computers | Chemical | Mechanical | Medical |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Patent scope | 0.951*** | 0.269 | 0.273 | 1.115** | 2.155* | 2.510*** | 9.115*** |
| | (0.255) | (0.206) | (0.636) | (0.468) | (1.117) | (0.775) | (3.295) |
| Observations | 339,769 | 53,368 | 123,404 | 30,199 | 43,990 | 41,605 | 7,458 |
| $R^2$ | 0.071 | 0.013 | 0.012 | 0.006 | 0.193 | 0.051 | 0.112 |
| Adjusted $R^2$ | 0.068 | 0.011 | 0.010 | -0.001 | 0.189 | 0.048 | 0.106 |
| Residual Std. Error | 11.476 | 10.579 | 11.600 | 9.724 | 12.327 | 12.172 | 17.375 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Covariate normalized as a z-score. All models run with application year by art unit fixed effects. Standard errors are clustered at the firm level.
A one standard deviation increase in examiner score increases patent scope by about .2 standard deviations.

Table 24 repeats the main results controlling for firm size. Because patents are randomly assigned to patent examiners and because firm size is demeaned, the main effect of patent scope on the likelihood that a patent is sold is not appreciably altered by including firm size as a covariate. The results show that in the unrestricted sample, patents by larger firms are less likely to be sold—a 100% increase in firm size decreases the likelihood that a patent is sold by .27 percentage points (p<.01). Further, the effect of an increase in patent scope on the likelihood of a sale decreases with firm size–a 100% increase in firm size decreases the effect of patent scope by .15 percentage points (p<.05). However, on the industry level the negative interaction is significant only for the computer and mechanical subsamples. Indeed, for chemical patents the effect is reversed—scope has a greater effect on patent sales for larger companies. Together these results provide qualified support for Hypothesis 5, suggesting that patent scope is in general more important for facilitating sales by smaller firms but that the relationship between patent strategy and firm size is industry-specific.

Table 25 presents estimates of conditional logistic regressions corresponding to those in Table 22, with the logit of post-grant sale regressed directly on examiner score. The coefficients are presented as log odds, but to aid in interpretability Figure 29 plots the probability of a sale for different industries as curves predicted based on these estimates. The results are broadly consistent with those shown in Tables 22 and 23. In the unrestricted sample, a one standard deviation increase in examiner score increases the odds that the patent is sold by a factor of 1.22 (p<.01). The effect is larger and statistically significant in the computer, chemical, and mechanical industries and largest in the medical device

**Table 24: Instrumental variable results by industry with firm size interactions.**

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Likelihood of sale (percentage points) | | | | | | |
| | All | Electrical | Semicond. | Computers | Chemical | Mechanical | Medical |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Patent scope | 0.949*** | 0.281 | 0.437 | 1.503** | 2.480** | 2.166*** | 9.014*** |
| | (0.255) | (0.230) | (0.705) | (0.617) | (1.175) | (0.639) | (3.298) |
| Firm size | −0.272*** | −0.317*** | −0.304*** | −0.350*** | −0.087 | −0.226*** | −0.391** |
| | (0.019) | (0.050) | (0.031) | (0.039) | (0.070) | (0.039) | (0.172) |
| Scope * size | −0.145** | −0.025 | −0.299 | −0.352** | 0.481* | −0.351* | 0.096 |
| | (0.071) | (0.072) | (0.217) | (0.139) | (0.256) | (0.203) | (0.815) |
| Observations | 339,769 | 53,368 | 123,404 | 30,199 | 43,990 | 41,605 | 7,458 |
| $R^2$ | 0.074 | 0.021 | 0.014 | 0.007 | 0.155 | 0.046 | 0.127 |
| Adjusted $R^2$ | 0.071 | 0.019 | 0.012 | -0.00005 | 0.151 | 0.043 | 0.121 |
| Residual Std. Error | 11.455 | 10.536 | 11.589 | 9.719 | 12.614 | 12.203 | 17.229 |

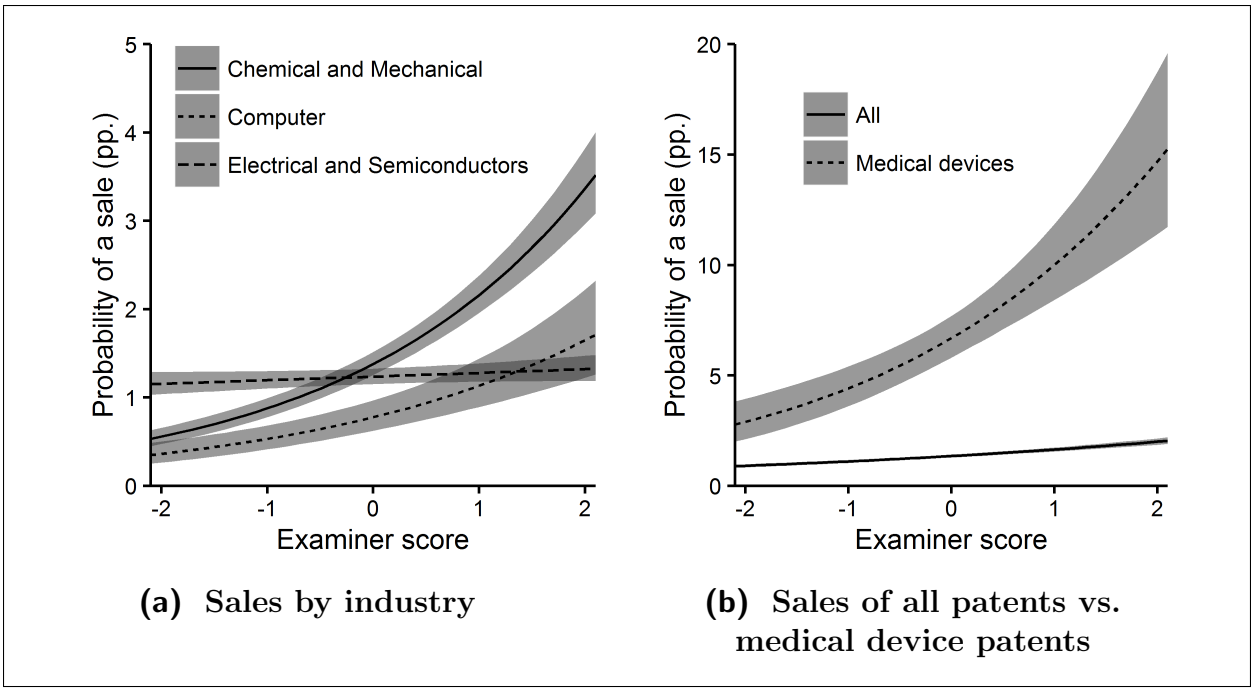
*Note:* *p<0.1; **p<0.05; ***p<0.01
Patent scope normalized as a z-score. All models run with application year by art unit fixed effects. Standard errors are clustered at the firm level.
Firm size operationalized as the log of patents issued in the past two years.
A one standard deviation increase in examiner score increases patent scope by about .2 standard deviations.

industry, where a one standard deviation increase in examiner score increases the odds that the patent is sold by a factor of 2.28 (p<.01). As is shown in Figure 29(a), industries that are largely similar in the overall likelihood that a patent is sold differ dramatically in the effect of examiner score on the likelihood of a sale. Figure 29(b), which compares the effect of examiner score in the medical device industry to the effect for all patents, highlights the outsized role that the right to exclude plays in certain industries.

**Table 25: Main results by industry, logistic regression.**

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Logit of sale | | | | | | |
| | All | Electrical | Semicond. | Computers | Chemical | Mechanical | Medical |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Examiner score | 0.200*** | 0.067 | 0.031 | 0.261*** | 0.215*** | 0.307*** | 0.824*** |
| | (0.015) | (0.042) | (0.027) | (0.061) | (0.044) | (0.046) | (0.106) |
| Observations | 339,769 | 53,368 | 123,404 | 30,199 | 43,990 | 41,605 | 7,458 |
| Log Likelihood | -25,080.010 | -3,108.598 | -8,378.250 | -1,455.206 | -3,018.093 | -2,872.838 | -648.316 |
| Akaike Inf. Crit. | 50,198.010 | 6,437.195 | 17,130.500 | 3,318.412 | 6,446.185 | 6,013.676 | 1,400.632 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
Covariate normalized as a z-score. All models run with application year by art unit fixed effects.




**(a) Sales by industry**

**(b) Sales of all patents vs. medical device patents**

**Figure 29: Probability that a patent is sold as a function of examiner score, by industry, predicted based on logistic regression estimates.**

In order to address concerns related to possible violation of the exclusion restriction for the instrumental variable and concerns about model fit with a low-probability binary dependent variable, the remainder of the results tables estimate reduced form conditional logistic regression models rather than two-stage least squares regression models. However, two-stage least squares regression tables corresponding with the tables in Section 2.4.2 are included in Appendix A for robustness.

### Industry and Market Characteristics

Although the effect of patent scope exhibits considerable heterogeneity across industry divisions, this cross-classifications does not precisely reflect reports of where and how market participants believe that patents are effective. Table 26 and Table 27 present the results of eight models examining how the effect of patent scope on the likelihood of a sale interacts with the effectiveness of patents in different industries. As discussed in Appendix B.1, survey measures of patent effectiveness provided by Cohen et al. (2000) were converted to z-scores by matching industries to art units and normalizing the measures across the sample of patents. In each table, models 1 and 2 present the results for product patent effectiveness, while models 3 and 4 present the results for process patent effectiveness. The coefficients are presented as log odds, but to aid in interpretation Figure 30(a) and Figure 30(b) graphically present the estimates of model 1 in Table 26 and Table 27.

The models presented in Table 26 include interaction terms for overall patent effective-

ness (models 1 and 3) and patent effectiveness for market exclusion (models 2 and 4).[2] Table 26 shows that the effect of examiner score on the likelihood of a sale increases as the reported effectiveness of patents increases overall (models 1 and 3) and specifically for exclusion (models 2 and 4). Indeed, these interaction effects are in most instances (models 1, 2, and 4) larger than the main effects, which are even negative for exclusion (models 2 and 4). As shown in Figure 30(a), in industries where product patents are reported as being ineffective for market exclusion, patents that vary widely in terms of their scope differ little in terms of the probability that they will be sold. In contrast, in industries where patents are reported as being effective for market exclusion, the probability that a patent with an examiner score one standard deviation greater than the mean is sold is over twice that for a patent with an examiner score one standard deviation narrower than the mean. Thus, patent effectiveness in general and for exclusion in particular as reported by industry participants may in part reflect the extent to which a patent's scope provides market safety for the purpose of facilitating sales.

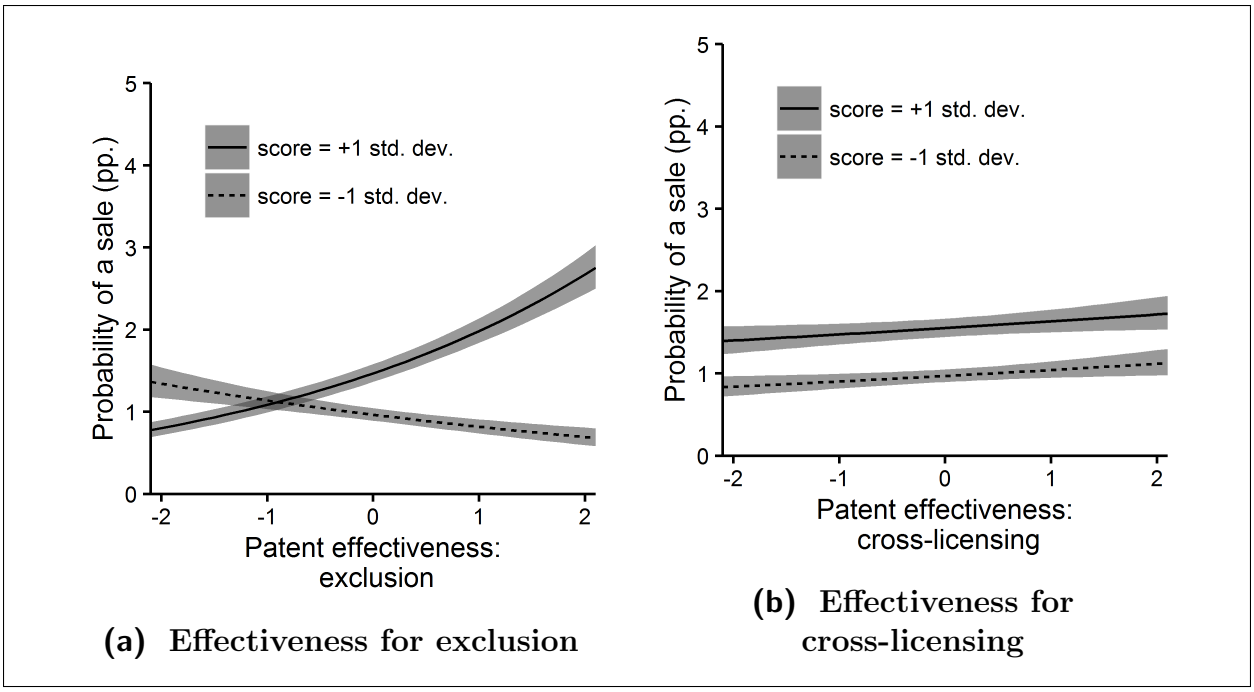
**Table 26: Logistic regressions by transfer type and industry-reported patent effectiveness.**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Logit of sale | | | |
| | Product | | Process | |
| | (1) | (2) | (3) | (4) |
| Examiner score | 0.221*** | 0.206*** | 0.227*** | 0.231*** |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| CNW Overall | 0.163*** | | 0.347*** | |
| | (0.023) | | (0.024) | |
| CNW Exclusion | | −0.026 | | −0.254*** |
| | | (0.031) | | (0.029) |
| Score x CNW Overall | 0.177*** | | 0.119*** | |
| | (0.018) | | (0.019) | |
| Score x CNW Exclusion | | 0.256*** | | 0.237*** |
| | | (0.019) | | (0.020) |
| Observations | 207,037 | 207,037 | 207,037 | 207,037 |
| $R^2$ | 0.031 | 0.031 | 0.035 | 0.031 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All covariates normalized as z-scores. All models run with application year by technology center fixed effects.

One concern with these results might be that any increase in patent importance or effectiveness, even for non-exclusionary purposes, may be associated with an increased effect

[2]The categories "Overall" and "Exclusion" correspond to the labels "Patents" and "Fences" in Cohen et al. (2000).




**(a) Effectiveness for exclusion**

**(b) Effectiveness for cross-licensing**

**Figure 30: Probability that a patent is sold as a function of industry-reported product patent effectiveness described by Cohen, Nelson & Walsh (2000), by examiner score, predicted based on logistic regression estimates.**

of patent scope on the likelihood of a sale. Table 27 reveals that this is not the case. Using patents for cross-licensing or negotiating with rivals reflects a portfolio-based strategy in which a single patent is less likely to be strategically important. In such a context, the scope of a patent is unlikely to have a particularly strong effect on the likelihood of a sale. The models presented in Table 27 include interaction terms for patent effectiveness for cross-licensing (models 1 and 3), and patent effectiveness for negotiating with rivals (models 2 and 4).[3] The main effects for each of these models are positive—an increase in patent effectiveness for negotiation or cross-licensing is associated with an increased likelihood that a given patent is sold. However, the effect of patent scope on the likelihood of a sale does not increase with the reported effectiveness of product patents for cross-licensing (model 1) or negotiation (model 3). In fact, the effect decreases as the reported effectiveness of process patents for cross-licensing (model 2) or negotiation (model 4) increases.

[3]The categories "Cross-licensing" and "Negotiation" correspond to the labels "Cross-licensing" and "Player" in Cohen et al. (2000).

**Table 27: Logistic regressions by transfer type and industry-reported patent effectiveness.**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Logit of sale | | | |
| | Product | | Process | |
| | (1) | (2) | (3) | (4) |
| Examiner score | 0.249*** | 0.248*** | 0.259*** | 0.259*** |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| CNW Cross-licensing | 0.195*** | | 0.331*** | |
| | (0.026) | | (0.024) | |
| CNW Negotiation | | 0.188*** | | 0.211*** |
| | | (0.027) | | (0.023) |
| Score x CNW Cross-licensing | −0.007 | | −0.073*** | |
| | (0.019) | | (0.019) | |
| Score x CNW Negotiation | | −0.013 | | −0.128*** |
| | | (0.019) | | (0.019) |
| Observations | 207,037 | 207,037 | 207,037 | 207,037 |
| $R^2$ | 0.027 | 0.026 | 0.031 | 0.029 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All covariates normalized as z-scores. All models run with application year by technology center fixed effects.

Overall, the results presented in Table 26 and Table 27 are consistent with Hypothesis 2. The likelihood that a patent is sold increases not only with the patent's scope, but also with the interaction of patent scope with the industry-reported effectiveness of the legal right to exclude. In contrast, the likelihood that a patent is sold does not increase with the interaction between the scope of the right to exclude and the effectiveness of patents for negotiation and cross-licensing.

The realized effect of patent scope on the likelihood of a sale thus reflects industry reports about where patents are effective for exclusion as opposed to more aggregate purposes such as cross-licensing or negotiation. While this differential effectiveness may in part reflect fundamental aspects of the underlying technology, it may also relate to characteristics of the market for technology. Table 28 presents the results of nine models examining how market characteristics moderate the relationship between the examiner score and different types of transfers. The dependent variable in the first three models is the logit of whether the patent is sold, and all coefficients are statistically significant for these models with p<.01. In model 2, the examiner score remains a positive and statistically significant predictor of a sale even when controlling for market thickness and the interaction of scope and thickness. Moreover, the effect of examiner score increases substantially in thicker markets. Figure 31(a) presents this result graphically by plotting the probability of a sale as a function of market thickness for patents one standard deviation above and below the mean examiner score. As is shown in Figure 32(a), broad and narrow patents have nearly the same probability of being sold in very thin markets, but broader patents are over three times more likely to be sold in very

thick markets. Model 3 presents a similar result for patent congestion. Patents are much less likely to be sold in more congested markets, and the effect of patent scope on the likelihood of a sale is substantially reduced in congested markets. Figure 33(b) presents this result graphically. Again, broad and narrow patents have nearly the same probability of being sold in very congested markets, but broader patents are much more likely to be sold in very uncongested markets. The results presented in Table 28 are thus broadly consistent with Hypotheses 3 and 4.

**Table 28: Logistic regressions by transfer type and market characteristics.**

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Logit of sale | | | Logit of exit | | | Logit of spinoff | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Examiner score | 0.200*** | 0.177*** | 0.189*** | 0.219*** | 0.184*** | 0.118*** | 0.121*** | 0.126*** | 0.121*** |
| | (0.015) | (0.017) | (0.016) | (0.027) | (0.028) | (0.033) | (0.022) | (0.023) | (0.022) |
| Market thickness | | 0.454*** | | | 0.291*** | | | 0.330*** | |
| | | (0.016) | | | (0.028) | | | (0.022) | |
| Patent congestion | | | −0.298*** | | | −0.832*** | | | −0.174*** |
| | | | (0.017) | | | (0.026) | | | (0.024) |
| Score x thickness | | 0.062*** | | | 0.122*** | | | −0.022 | |
| | | (0.016) | | | (0.027) | | | (0.022) | |
| Score x congestion | | | −0.070*** | | | −0.153*** | | | −0.023 |
| | | | (0.016) | | | (0.025) | | | (0.023) |
| Observations | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 |
| $R^2$ | 0.018 | 0.037 | 0.026 | 0.053 | 0.061 | 0.105 | 0.015 | 0.023 | 0.017 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All covariates normalized as z-scores. All models run with application year by technology center fixed effects.

The models presented in Table 28 include fixed effects for broad industry categories. Nevertheless, one concern with these results is that market characteristics are highly endogenous, so an unobserved variable correlated with market characteristics rather than the market characteristics themselves may be responsible for the change in the effect of patent scope. For instance, industry structure may both give rise to market characteristics and influence the effect of patent scope. This concern cannot be completely ruled out, but one way of addressing it is to restrict the sample to a specific industry that nevertheless contains subindustries that vary in market characteristics. The medical device industry works for this purpose since it is composed of a number of different technologies that vary in their patent market characteristics, with large firms buying patents from different areas of medical device technology. Figure 31(c) and Figure 31(d) plot the effect of plus or minus one standard deviation in examiner score on the likelihood of a sale as a function of market thickness and market congestion for the subsample patents in the medical device industry, with the x-axis limited to where the data provides full support. The interactions between examiner score and market characteristics for the medical device subsample are consistent with the results

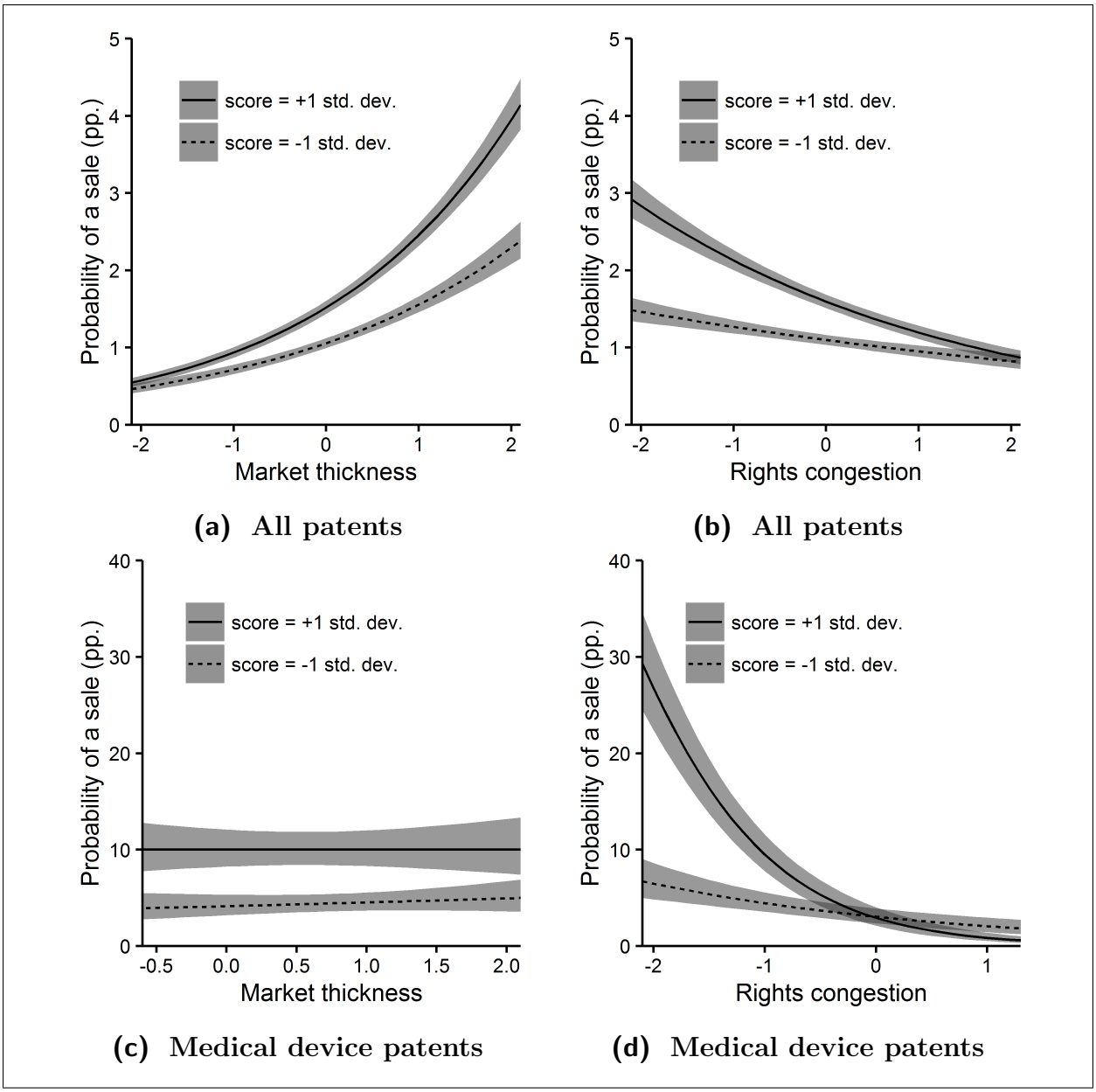



**Figure 31: Probability that a patent is sold as a function of market characteristics, by examiner score, predicted based on logistic regression estimates.**

for the unrestricted sample, with an increase in examiner score having virtually no effect in thin or congested markets and a very large effect in thick or uncongested markets.

Patents are transferred not only in sales between firms, but also via firm exits (e.g., bankruptcy, acquisition, merger) and firm spinoffs. A patent transfer via a firm exit may be in many cases a form of market transaction—this is clear in the case of bankruptcy, but even in mergers and acquisitions a large part of a firm's value may depend on its patent portfolio. Thus, patent transfers via firm exits should exhibit many of the same characteristics as patent

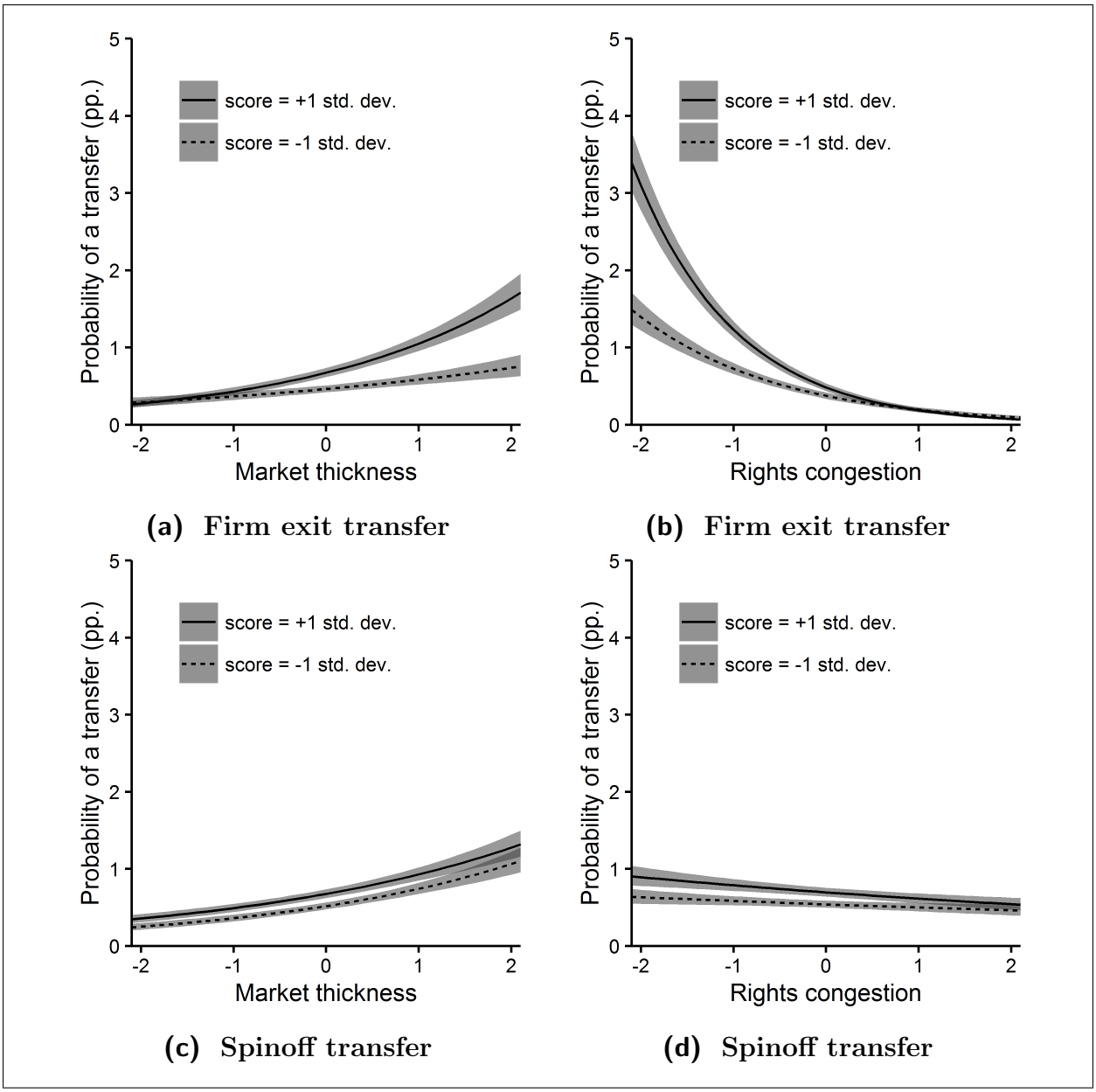



**Figure 32: Probability that a patent is transferred via a firm exit or a spinoff as a function of market characteristics, by examiner score, predicted based on logistic regression estimates.**

sales. Models 4, 5, and 6 in Table 26, in which the dependent variable is the logit of whether the patent is transferred in a firm exit from the patent market, are consistent with this story. Across each of these models, all coefficients are statistically significant with $p<.01$. Examiner score has a positive effect on the logit of post-grant exit transfers. Examiner score interacts positively with market thickness and negatively with market congestion. Figure 32(a) and Figure 32(b) present the results of models 5 and 6 graphically. As shown in Figure 32(a), in a very thick market the probability of a firm exit transfer is over three times greater for a patent

having a score one standard deviation greater than the mean than for a patent having an examiner score one standard deviation narrower than the mean, while the difference between high and low examiner scores is negligible in a thin market. Similarly, Figure 32(b) shows that patents in congested markets are almost never transferred via firm exits, and that the examiner score has little effect on the likelihood of such a transfer. However, the examiner score exerts a large and positive effect on firm exit transfers in very uncongested markets.

In contrast to firm exits, spinoff transfers are not conducted in a marketplace. Because a spun off firm typically has a close relationship to the parent firm and because the spinoff is unlikely to involve a sharp negotiation over patent rights, the likelihood that a patent is transferred via a spinoff should not depend very much on the scope of the patent. Models 7, 32, and 9 in Table 26, in which the dependent variable is the logit of whether the patent is transferred in a firm spinoff, are consistent with this story. An increase in examiner score slightly but significantly increases the likelihood of a spinoff transfer—a one standard deviation increase in examiner score increases the odds of such a transfer by a factor of 1.13 ($p<.01$). However, the interactions between examiner score and market characteristics are small in magnitude and insignificant. Figure 32(c) and Figure 32(d) present the results of models 8 and 9 graphically.

**Conclusion**

The results demonstrate that an exogenous increase in the scope of a patent's legal right to exclude, holding fixed the underlying idea, increases the likelihood that the patent is sold. A patent having a broader exclusionary right is likely more valuable, so the sale of that patent is more likely to overcome the considerable transaction costs involved in the transfer of intellectual property. The effect is stronger for smaller firms and in general holds not only for patent sales but also for the transfer of patents upon firm exit. Broadly speaking, this finding highlights the importance of the legal right to exclude as a strategic factor for firms seeking to profit from innovation by transferring intellectual property rights. (Gans et al., 2008; Teece, 1986).

The results also show that the effect of patent scope on the likelihood that the patent is sold varies considerably across industries. The effect is large in the medical device, chemical, mechanical, and computer industries, but is small and not statistically significant in the electronic component and semiconductor industries. The results suggest that these differences may reflect in part differences in the effectiveness of the legal right to exclude for providing market exclusivity, which at the industry level is distinct from both the overall rate of patent sales and the effectiveness of patents for purposes other than exclusion. (Cohen et al., 2000).s For example, patents are no more likely to be sold in industries where they are reported as being more effective for exclusion, but the effect of patent scope on the likelihood of a sale increases with the effectiveness of patents for exclusion. However, although patents are more likely to be sold in industries where they are more effective for cross-licensing, the effect of patent scope on the likelihood of a sale is unrelated to the effectiveness of patents for cross-licensing.

At a more detailed level, the results highlight how the relationship between patent scope and patent sales depends substantially on characteristics of the market in which the patent is situated, even within the broad industry categories noted above. Viewed through the lens

of the market design literature (see, e.g., Roth, 2008) and consistent with Arrow (1962), the legal right to exclude provides market safety by ensuring that the purchaser of an idea actually receives ownership of the idea being purchased. As in all markets, the results suggest that patent sales are inhibited by congestion, a lack of market safety, and a lack of market thickness. (Roth (2008)). In a finding specific to patent markets, the results show that patent scope has less of an effect on the likelihood of a sale in a thin or congested market, suggesting that patent scope is less effective at providing market safety in these contexts. In particular, congested patent rights seem to render individual patents less effective, consistent with the findings of Ziedonis (2004) and Hall et al. (2015). Together these demonstrate that just as the market for ideas through patent licensing is plagued with imperfections (see, e.g., Agrawal et al., 2015; Arora et al., 2004; Gans & Stern, 2010), so too is the free flow of rights on the secondary market for patents.

Although only about 5.7% of patents are transferred post-grant, the data suggests that many of these transfers are strategic and important. Patents that are transferred are more highly cited and are more likely to block other patents via novelty or non-obviousness rejections. Patent transfers also seem to reflect choices by large firms about firm boundaries, since such firms are more likely to transfer patents on technology that is unfamiliar to them and less likely to transfer patents on technology that is more original. The secondary market for patents may thus be important for a range of strategic considerations such as establishing a technological competitive advantage (Teece, 1982), reducing IP ownership fragmentation (Ziedonis, 2004), and allowing a firm to profit from innovation that it is poorly equipped to commercialize (Teece, 1986). These strategic considerations may explain in part the economic growth and gains to trade that recent studies have found to be associated with patent sales. (Akcigit et al., 2016; Serrano, 2011). In addition, strategic transfers may spur further development of the underlying technology rather than allowing it to languish unexploited, for instance when technologies radical to the firm are orphaned when an intrapreneur leaves the organization. (See, e,g. Pinchot, 1987).

The results also highlight the complex tradeoffs inherent in deciding issues related to patent grant rates and patent scope determinations. (Merges & Nelson, 1990). The promise of a patent is intended to provide incentives for innovation, and ownership of a patent can also provide incentives for further development and commercialization. Indeed, awarding broad rights for a particular patent may facilitate the patent's transfer to a firm better able to develop the underlying ideas. However, consistent with the findings of Ziedonis (2004), a surfeit of overlapping and uncertain rights can create market congestion that inhibits the free flow of rights between firms.

I argue that these results supports three main conclusions. First, holding fixed the underlying idea, a broader patent is more likely to be sold. Second, patent sales often represent strategic decisions for firms that stem in part from the strategic importance of the underlying ideas, in part from the legal characteristics of the intellectual property right, and in part from the market characteristics of the technology in which the patent is situated. Third, the effect of the scope of the legal right to exclude on firm decisions differs substantially across firms, markets, and industries, highlighting both variation in the effectiveness of patent rights and the importance of within-industry analysis for better understanding the strategic implications of the patent system for firms. Together these conclusions suggest that both patent scope and patent sales are important components of the market for ideas and may play an

important role in determining why some firms succeed while others fail.

# Chapter 5

## Conclusion

This thesis provides an integrated exploration of the role of the scope of a patent's legal right to exclude in firm strategy and innovation economics. Chapter 2 introduces a new measure of patent-to-patent technological similarity, which is used to support the empirical analysis in Chapter 3 and Chapter 4. Chapter 3 examines one of the causes of receiving broader patent scope—winning a patent race—while Chapter 4 investigates one of the consequences of receiving broader patent scope—a lower likelihood that a patent is sold.

Chapter 2 of this thesis, co-authored with Kenneth Younge, provides descriptive evidence in support of scholars arguments that existing measures of technological space are too coarse for many research questions (Thompson and Fox-Kean 2005). Further, it introduced a new measure of patent-to-patent similarity based on a cosine distance measure applied to a tf-idf weighting of the full text of every pair of more than 14 trillion pairs of patents. The resulting measure provides a single, simple, continuous, standardized, bounded value for patent-to-patent similarity that is free of selection effects and that serves as an alternative to a variety of previous approaches based on manual classification.

We demonstrate the validity of the measuring using a range of techniques. The measure is based on recent theoretical and empirical work in information retrieval theory. It correlates well with inventor, patent attorney, and lay person notions of relatedness; it predicts patent citations better than existing class-based measures; and it matches with patent phenomena both internal (e.g., patent families) and external (e.g., firm technological focus) to the patent system.

In follow-on work, we plan to make the patent-to-patent similarity measure available to scholars and to apply the measure to study various phenomena of interest to economists and strategy scholars. For example, the patent-to-patent similarity measure may be used to map the relative positioning of firms in a unified technological space (e.g., Stuart and Podolny 1996), compare spillovers and rivalry in both technology and product market space (e.g., Bloom et al. 2013), detect technology entry and portfolio competition (e.g., Hall and Ziedonis 2001), measure the relatedness and overlap of alliance partners (e.g., Mowery et al. 1996), construct measures of prior activity and breakthroughs in a space (e.g., Ahuja and Lampert 2001), evaluate the impact of venture capital on subsequent innovation (e.g., Kortum and Lerner 2000), examine the impact of patent scope (e.g., Merges and Nelson 1990), identify patent thickets and situations of holdup (e.g., Lemley and Shapiro 2006), track the technological trajectory of patent races (e.g., Lerner 1997), control for the preexisting distribution of inventive activity (e.g., Jaffe et al. 1993), and assess the quality and relevance of patent citations. (Trajtenberg 1990).

Chapter 3 of this thesis, co-authored with Neil Thompson, provides the first large sample empirical analysis of patent races. We use detail administrative records from patent examiners to develop a novel approach for identifying patent races. We find that patent racing is common, with 10-11% of all patents being in races and with large firms being more likely to race. We also find that racing is much more common in areas of technology thought to have weaker patents (e.g., computers and communications). Finally, we show that patents in

races (both leaders and followers) are much more likely to have their scope narrowed during examination than patents not in races.

Next, we use a regression discontinuity approach to estimate the causal effects of winning and losing a patent race. Race leaders tend to receive broader patents than followers, and are also more likely to pay to maintain their patents throughout the patent term. However, about 67% of race followers still receive patent coverage from their discovery. Most importantly, we show that winning a patent race affects follow-on innovation, both in direction and magnitude. Leaders conduct significantly more follow-on research, and the work that they do is closer to the original patent than is follow-on research by followers. In follow-on work, we plan to investigate the implications of these findings for theoretical models of patent racing.

Chapter 4 of this thesis examines the effect of patent scope on patent sales. The results show that an exogenous increase in the scope of a patent's legal right to exclude, holding fixed the underlying idea, increases the likelihood that the patent is sold. Broadly speaking, this finding highlights the importance of the legal right to exclude as a strategic factor for firms seeking to profit from innovation by transferring intellectual property rights. (Gans et al. 2008, Teece 1986). However, the magnitude of the effect depends considerably on endogenous firm and market characteristics. In particular, broader patent scope exerts a larger effect on the sale of patents by small firms and the sale of patents in thick and uncongested markets, results which are consistent with the literature on transaction cost economics (Arrow 1962) and market design (Roth 2008). Broader patent scope also exerts a larger effect in industries where patents are reported as being effective for exclusion, results which are consistent with the literature on the differential role patents play across different industries. (Cohen et al. 2000).

In follow-on work, I plan to focus the analysis on small firms, on which the patent system is thought to exert a disproportionate impact. I also plan to extend the analysis to study the purchase side of patent sales. For example, a study of the purchase side of patent sales would focus attention on the choices of firms that expand their technological capabilities in part by purchasing the rights to technology, which can provide efficieny gains by helping to avoid duplicative research investments by different firms. However, any efficiency gains brought about by increased sales caused by a reduction in transaction costs may be more than offset if broad patents are primarily purchased by non-practicing entities that extract money from firms through the threat of litigation.

This thesis presents an in-depth exploration of the consequences of the patent system by conducting large-sample empirical analysis of data sets constructed by analyzing the full text, claims, and metadata of millions of patents and patent applications. Together the results supports the argument that whether a firm has a patent is often less important than the characteristics of that patent. For example, the results demonstrate that even when losing a patent races, a firm still typically receives a patent. However, the winner of a race typically receives a much broader patent than the loser and is more likely to conduct related follow-on research. Further, the results demonstrate that a broader patent is more likely to be sold. However, the effect depends on firm and market characteristics. These results are made possible in part by the development of patent-to-patent textual similarity data, which provides the raw materials for generating measures of patent impact, novelty, and familiarity, among others.

# Bibliography

Agarwal, R., M. Ganco and R. H. Ziedonis (2009). "Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility." Strategic Management Journal 30(13): 1349-1374.

Agrawal, A., Cockburn, I., & Zhang, L. (2015). Deals not done: Sources of failure in the market for ideas. Strategic management journal, 36(7), 976-986.

Ahuja, G. (2000). "Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study." Administrative Science Quarterly 45(3): 425-455.

Ahuja, G. and C. M. Lampert (2001). "Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions." Strategic Management Journal 22(6-7): 521-543.

Aizawa, A. (2003). "An information-theoretic perspective of tf–idf measures." Information Processing & Management 39(1): 45-65.

Akcigit, U., Celik, M. A., & Greenwood, J. (2016). Buy, Keep, or Sell: Economic Growth and the Market for Ideas. Econometrica, 84(3), 943-984.

Allison, J. R. and M. A. Lemley (2002). "Growing Complexity of the United States Patent System, The." BUL Rev. 82: 77.

Almeida, P. and B. Kogut (1999). "Localization of Knowledge and the Mobility of Engineers in Regional Networks." Management Science 45(7): 905-917.

Anand, B. N., & Khanna, T. (2000). The structure of licensing contracts. The Journal of Industrial Economics, 48(1), 103-135.

Anton, J. J., & Yao, D. A. (2002). The sale of ideas: Strategic disclosure, property rights, and contracting. The Review of Economic Studies, 69(3), 513-531.

Arora, A., & Ceccagnoli, M. (2006). Patent protection, complementary assets, and firms' incentives for technology licensing. Management Science, 52(2), 293-308.

Arora, A., & Gambardella, A. (2010). Ideas for rent: an overview of markets for technology. Industrial and Corporate Change, 19(3), 775-803.

Arora, A., Belenzon, S., & Rios, L. A. (2011). The organization of R&D in American corporations: the determinants and consequences of decentralization. National Bureau of Economic Research.

Arora, A., Fosfuri, A., & Gambardella, A. (2004). Markets for technology: The economics of innovation and corporate strategy: MIT press.

Arrow, K. (1962). Economic Welfare and the Allocation of Resources for Invention. The Rate and Direction of Inventive Activity: Economic and Social Factors. R. Nelson. Princeton, NJ, Princeton University Press: 609-626.

Aspex Eyewear, Inc. v. Miracle Optics, Inc, No. No. 04-1265, 434 1336 (Court of Appeals, Federal Circuit 2006).

Bar, T. and A. Leiponen (2012). "A measure of technological distance." Economics Letters 116(3): 457-459.

Belenzon, S. and M. Schankerman (2013). "Spreading the word: Geography, policy, and knowledge spillovers." Review of Economics and Statistics 95(3): 884-903.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, "The lifecycle of inventors," (mimeo Harvard, 2015).

Benner, M. and J. Waldfogel (2008). "Close to you? Bias and precision in patent-based measures of technological proximity." Research Policy 37(9): 1556-1567.

Bessen, J. E., Meurer, M. J., & Ford, J. L. (2011). The private and social costs of patent trolls. Boston Univ. School of Law, Law and Economics Research Paper(11-45).

Bikard, Michaël, "Peer-Based Knowledge Validation: A Hurdle to the Flow of Academic Science to Inventors," Available at SSRN 2333413, (2015).

Bloom, N., M. Schankerman and J. Van Reenen (2013). "Identifying technological spillovers and product market rivalry." Econometrica 81(4): 1347-1393.

Burk, D. L., & Lemley, M. A. (2003). Policy levers in patent law. Virginia Law Review, 1575-1696.

Chen, Y.-L. and Y.-T. Chiu (2011). "An IPC-based vector space model for patent retrieval." Information Processing & Management 47(3): 309-322.

Coase, R. H. (1937). The nature of the firm. economica, 4(16), 386-405.

Cockburn, I. M., Kortum, S., & Stern, S. (2002). Are all patent examiners equal?: The impact of characteristics on patent statistics and litigation outcomes: National Bureau of Economic Research.

Cockburn, I. M., MacGarvie, M. J., & Mueller, E. (2010). Patent thickets, licensing and innovative performance. Industrial and Corporate Change.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not). Retrieved from

Cohen, Wesley M, Richard R Nelson, and John P Walsh, "Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not)," (National Bureau of Economic Research, 2000).

Cortina, J. (1993). "What Is Coefficient Alpha? An Examination of Theory and Applications." Journal of Applied Psychology 78(1): 98-104.

Cronbach, L. (1971). Test validation. Educational measurement. R. L. Thorndike. Washington, DC, American Council of Education 443-507.

Crouch, Dennis D, "Is Novelty Obsolete? Chronicling the Irrelevance of the Invention Date in US Patent Law," Michigan Telecommunications and Technology Law Review, 16 (2010).

Dahlman, C. J. (1979). The problem of externality. The Journal of Law & Economics, 22(1), 141-162.

Dasgupta, Partha, and Joseph Stiglitz, "Industrial structure and the nature of innovative activity," The Economic Journal, (1980), 266-293.

Dasgupta, Partha, and Joseph Stiglitz, "Uncertainty, industrial structure, and the speed of R&D," The Bell Journal of Economics, (1980), 1-28.

Doraszelski, Ulrich, "An R&D race with knowledge accumulation," RAND Journal of economics, (2003), 20-42.

Elkan, C. (2005). Deriving TF-IDF as a Fisher Kernel. String Processing and Information Retrieval. M. Consens and G. Navarro, Springer Berlin Heidelberg. 3772: 295-300.

Fan, J., F. Han and H. Liu (2014). "Challenges of Big Data analysis." National Science Review 1(2): 293-314.

Fischer, T., & Henkel, J. (2012). Patent trolls on markets for technology–An empirical analysis of NPEs' patent acquisitions. Research Policy, 41(9), 1519-1533.

Fleming, L. (2001). "Recombinant Uncertainty in Technological Search." Management Science 47(1): 117-132.

Fosfuri, A. (2006). The licensing dilemma: understanding the determinants of the rate of technology licensing. Strategic management journal, 27(12), 1141-1158.

Frakes, M. D. and M. F. Wasserman (2014). "The Failed Promise of User Fees: Empirical Evidence from the U.S. Patent and Trademark Office." Journal of Empirical Legal Studies 11(4): 602-636.

Fudenberg, Drew, Richard Gilbert, Joseph Stiglitz, and Jean Tirole, "Preemption, leapfrogging and competition in patent races," European Economic Review, 22 (1983), 3-31.

Fuxman, A., P. Tsaparas, K. Achan and R. Agrawal (2008). Using the wisdom of the crowds for keyword generation. Proceedings of the 17th international conference on World Wide Web, ACM.

Galasso, A., & Schankerman, M. (2008). Patent thickets and the market for ideas: evidence from settlement of patent disputes.

Galasso, A., Schankerman, M., & Serrano, C. J. (2013). Trading and enforcing patent rights. The Rand Journal of Economics, 44(2), 275-312.

Gallini, N. T., & Wright, B. D. (1990). Technology transfer under asymmetric information. The Rand Journal of Economics, 147-160.

Gambardella, A., Giuri, P., & Luzzi, A. (2007). The market for patents in Europe. Research Policy, 36(8), 1163-1183.

Gambardella, Alfonso, Dietmar Harhoff, and Bart Verspagen, "The value of European patents," European Management Review, 5 (2008), 69-84.

Gans, J. S., & Stern, S. (2003). The product market and the market for “ideas”: commercialization strategies for technology entrepreneurs. Research Policy, 32(2), 333-350.

Gans, J. S., & Stern, S. (2010). Is there a market for ideas? Industrial and Corporate Change.

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.

Gans, Joshua S, David H Hsu, and Scott Stern, "The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays," Management Science, 54 (2008), 982-997.

Gilbert, Richard J, and David MG Newbery, "Preemptive patenting and the persistence of monopoly," The American Economic Review, (1982), 514-526.

Gilsing, V., B. Nooteboom, W. Vanhaverbeke, G. Duysters and A. van den Oord (2008). "Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density." Research Policy 37(10): 1717-1731.

Graham, S. J. H. and D. Harhoff (2014). "Separating patent wheat from chaff: Would the US benefit from adopting patent post-grant review?" Research Policy 43(9): 1649-1659.

Graham, S., & Hegde, D. (2015). Disclosing patents' secrets. Science, 347(6219), 236-237.

Griliches, Z. (1979). "Issues in Assessing the Contribution of Research and Development to Productivity Growth." The Bell Journal of Economics 10(1): 92-116.

Hall, B. H. and R. H. Ziedonis (2001). "The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995." RAND Journal of Economics 32(1): 101-128.

Hall, B. H., A. B. Jaffe and M. Trajtenberg (2001). The NBER patent Citations Data File: Lessons Insights and Methodological Tools, NBER.

Hall, B. H., Helmers, C., & Von Graevenitz, G. (2015). Technology entry in the presence of patent thickets. National Bureau of Economic Research.

Harris, Christopher, and John Vickers, "Patent races and the persistence of monopoly," The Journal of Industrial Economics, (1985), 461-481.

Harris, Christopher, and John Vickers, "Perfect Equilibrium in a Model of a Race," The Review of Economic Studies, 52 (1985), 193-209.

Hegde, D. (2012). "Funding and performance at the US Patent and Trademark Office." Nature Biotechnology 30(2): 148-150.

Henderson, R. and I. Cockburn (1996). "Scale, Scope, and Spillovers: The Determinants of Research Productivity in Drug Discovery." The RAND Journal of Economics 27(1): 32-59.

Henderson, R., A. Jaffe and M. Trajtenberg (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment." American Economic Review 95(1): 461-464.

Hochberg, Y. V., Serrano, C. J., & Ziedonis, R. H. (2014). Patent collateral, investor commitment, and the market for venture lending. National Bureau of Economic Research.

Jaffe, A. B. (1986). "Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value." The American Economic Review 76(5): 984-1001.

Jaffe, A. B. (1989). "Characterizing the “technological position” of firms, with application to quantifying technological opportunity and research spillovers." Research Policy 18(2): 87-97.

Jaffe, A. B., M. Trajtenberg and R. Henderson (1993). "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." The Quarterly Journal of Economics 108(3): 577-598.

Judd, By Kenneth L, Karl Schmedders, and Şevin Yeltekin, "Optimal Rules for Patent Races," International Economic Review, 53 (2012), 23-52.

King, G. and L. Zeng (2001). "Logistic Regression in Rare Events Data." Political Analysis 9(2): 137-163.

Klepper, S., & Sleeper, S. (2005). Entry by spinoffs. Management Science, 51(8), 1291-1306.

Kortum, S. and J. Lerner (2000). "Assessing the contribution of venture capital to innovation." RAND journal of Economics: 674-692.

Kuhn, J. M., & Younge, K. A. (2016). Patent Citations: An Examination of the Data Generating Process.

Kuhn, Jeffrey M and Neil C Thompson, "Causal Inference on Patent Protection," (2016).

Kuhn, Jeffrey M, and Kenneth A Younge, "Patent Citations: An Examination of the Data Generating Process," (2016).

Kuhn, Jeffrey M, Kenneth A Younge, and Alan Marco, "Patent Citations: An Examination of the Data Generating Process," (2016).

Lamoreaux, N. R., & Sokoloff, K. L. (1999). Inventors, firms, and the market for technology in the late nineteenth and early twentieth centuries Learning by doing in markets, firms, and countries (pp. 19-60): University of Chicago Press.

Lamoreaux, N. R., & Sokoloff, K. L. (2001). Market trade in patents and the rise of a class of specialized inventors in the 19th-century United States. The American Economic Review, 91(2), 39-44.

Lee, Tom, and Louis L Wilde, "Market structure and innovation: a reformulation," The Quarterly Journal of Economics, (1980), 429-436.

Lemley, M. A. (2001). Rational ignorance at the patent office. Northwestern University Law Review, 95(4).

Lemley, M. A. and C. Shapiro (2006). "Patent holdup and royalty stacking." Tex. L. Rev. 85: 2163.

Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. Review of Economics and Statistics, 94(3), 817-827.

Lemley, M. A., & Sampat, B. N. (2008). Is the Patent Office a Rubber Stamp? Emory Law Journal, 58, 181.

Lemley, M. A., & Shapiro, C. (2005). Probabilistic patents. The Journal of Economic Perspectives, 19(2), 75-98.

Lerner, J. (1997). "An Empirical Exploration of a Technology Race." The RAND Journal of Economics 28(2): 228-247.

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

Loury, Glenn C, "Market structure and innovation," The Quarterly Journal of Economics, (1979), 395-410.

Manning, C. D., P. Raghavan and H. Schütze (2008). Introduction to information retrieval. Cambridge, U.K., Cambridge University Press.

McNamee, R. C. (2013). "Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example." Research Policy 42(4): 855-873.

Merges, R. P. (1999). As many as six impossible patents before breakfast: Property rights for business concepts and patent system reform. Berkeley Technology Law Journal, 577-615.

Merges, R. P. and R. R. Nelson (1990). "On the complex economics of patent scope." Columbia Law Review: 839-916.

Merges, R. P., & Nelson, R. R. (1994). On limiting or encouraging rivalry in technical progress: the effect of patent scope decisions. Journal of Economic Behavior & Organization, 25(1), 1-24.

Ming, Z. Y., K. Wang and T. S. Chua (2010). Vocabulary filtering for termweighting in archived question search. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 6118 LNAI: 383-390.

Mowery, D. C., J. E. Oxley and B. S. Silverman (1996). "Strategic alliances and interfirm knowledge transfer." Strategic Management Journal 17: 77-91.

Nanda, R., K. A. Younge and L. Fleming (2015). Innovation and Entrepreneurship in Renewable Energy. The Changing Frontier: Rethinking Science and Innovation Policy. A. B. Jaffe and B. Jones. Cambridge, MA, NBER/University of Chicago Press.

Nelson, R. R. and E. S. Phelps (1966). "Investment in humans, technological diffusion, and economic growth." The American Economic Review: 69-75.

Nesta, L. and P. P. Saviotti (2005). "Coherence of the Knowledge Base and the Firm's Innovative Performance: Evidence from the U.S. Pharmaceutical Industry." The Journal of Industrial Economics 53(1): 123-142.

Niioka, H. (2006). Patent auctions: business and investment strategy in IP commercialization. Journal of Intellectual Property Law & Practice, 1(11), 728-731.

Nunnally, J. C. and I. H. Berstein (1994). Psychometric Theory. New York, NY, McGraw-Hill.

Pinchot, G. (1987). Innovation through intrapreneuring. Research management, 30(2), 14-19.

Podolny, J. and T. Stuart (1995). "A role-based ecology of technological change." American Journal of Sociology: 1224-1260.

Quinn, Gene, "Change? Derviation May Feel a Lot Like Interference Practice," in IPWatchdog, (2012).

Quinn, Gene, "Reform Doing Away with Interference Proceedings & First to Invent," (2010).

Reinganum, Jennifer F, "A dynamic game of R and D: Patent protection and competitive behavior," Econometrica: Journal of the Econometric Society, 50 (1982), 671-688.

Reinganum, Jennifer F, "Dynamic games of innovation," Journal of Economic Theory, 25 (1981), 21-41.

Richardson, K., Oliver, E., & Costa, M. (2016). The 2015 Brokered Patent Market: A Good Year to be a Buyer.

Risch, M. (2011). Patent troll myths.

Romer, P. M. (1990). "Endogenous Technological Change." The Journal of Political Economy 98(5): S71-S102.

Rosenberg, N. (1974). "Science, invention and economic growth." The Economic Journal: 90-108.

Rosenkopf, L. and A. Nerkar (2001). "Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry." Strategic Management Journal 22(4): 287-306.

Rosenkopf, L. and P. Almeida (2003). "Overcoming Local Search Through Alliances and Mobility." Management Science 49(6): 751-766.

Roth, A. E. (2008). What have we learned from market design? The Economic Journal, 118(527), 285-310.

Salton, G., A. Wong and C.-S. Yang (1975). "A vector space model for automatic indexing." Communications of the ACM 18(11): 613-620.

Sampat, B., & Williams, H. L. (2015). How do patents affect follow-on innovation? Evidence from the human genome. National Bureau of Economic Research.

Schmookler, J. (1966). Invention and Economic Growth. Cambridge, Harvard University Press.

Serrano, C. J. (2010). The dynamics of the transfer and renewal of patents. The Rand Journal of Economics, 41(4), 686-708.

Serrano, C. J. (2011). Estimating the gains from trade in the market for innovation: Evidence from the transfer of patents. National Bureau of Economic Research.

Short, J. C., J. C. Broberg, C. C. Cogliser and K. H. Brigham (2010). "Construct Validation Using Computer-Aided Text Analysis (CATA)." Organizational Research Methods 13(2): 320-347.

Silverman, B. S. (1999). "Technological Resources and the Direction of Corporate Diversification: Toward an Integration of the Resource-Based View and Transaction Cost Economics." Management Science 45(8): 1109-1124.

Singh, J. and A. Agrawal (2011). "Recruiting for ideas: how firms exploit the prior inventions of new hires." Management Science 57(1): 129-150.

Somaya, D. (2002). Theoretical perspectives on patent strategy. Strategic management journal, 24(1).

Somaya, D. (2003). Strategic determinants of decisions not to settle patent litigation. Strategic management journal, 24(1), 17-38.

Somaya, D. (2012). Patent strategy and management an integrative review and research agenda. Journal of Management, 38(4), 1084-1114.

Somaya, D., Kim, Y., & Vonortas, N. S. (2011). Exclusivity in licensing alliances: using hostages to support technology commercialization. Strategic management journal, 32(2), 159-186.

Song, J., P. Almeida and G. Wu (2003). "Learning-by-Hiring: When Is Mobility More Likely to Facilitate Interfirm Knowledge Transfer?" Management Science 49(4): 351-365.

Stuart, T. E. and J. M. Podolny (1996). "Local Search and the Evolution of Technological Capabilities." Strategic Management Journal 17: 21-38.

Teece, D. J. (1977). Technology transfer by multinational firms: The resource cost of transferring technological know-how. The Economic Journal, 87(346), 242-261.

Teece, D. J. (1980). Economies of scope and the scope of the enterprise. Journal of Economic Behavior & Organization, 1(3), 223-247.

Teece, D. J. (1982). Towards an economic theory of the multiproduct firm. Journal of Economic Behavior & Organization, 3(1), 39-63.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. Research Policy, 15(6), 285-305.

Thompson, N. and J. Kuhn (2016). Patent Races in The Real World: They're Different Than Theory Suggests. Working Paper: Massachusetts Instititute of Technology/University of California Berkeley.

Thompson, N., & Kuhn, J. (2016). Does Winning a Patent Race Lead to More Follow-on Innovation?

Thompson, P. and M. Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." The American Economic Review 95(1): 450-460.

Traditional Total Pendency, Including RCEs, Oct. 2014 – Nov. 2016 (http://www.uspto.gov/corda/dashboards/patents/kpis/kpiWithRCE.kpixml,

Trajtenberg, M. (1990). "A Penny for Your Quotes: Patent Citations and the Value of Innovations." The RAND Journal of Economics 21(1): 172-187.

Turney, P. D. and P. Pantel (2010). "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research 37(1): 141-188.

Von Graevenitz, G., Wagner, S., & Harhoff, D. (2011). How to measure patent thickets—A novel approach. Economics letters, 111(1), 6-9.

Williamson, O. E. (1975). Markets and hierarchies. New York, 26-30.

Wu, H. C., R. W. P. Luk, K. F. Wong and K. L. Kwok (2008). "Interpreting TF-IDF term weights as making relevance decisions." ACM Trans. Inf. Syst. 26(3): 1-37.

Younge, K. A. and B. C. Römer (2016). The Geography of Knowledge Spillovers: A Reassessment Based on Patent-to-Patent Similarity. Working Paper: École Polytechnique Fédérale de Lausanne/Colorado State University.

Younge, Kenneth A, and Jeffrey M Kuhn, "Patent-to-Patent Similarity: A Vector Space Model," Available at SSRN, (2015).

Ziedonis, R. H. (2004). Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. Management Science, 50(6), 804-820.

Zizzo, Daniel John, "Racing with uncertainty: a patent race experiment," International Journal of Industrial Organization, 20 (2002), 877-902.

**Appendix A: Robustness checks**

Tables A.1-A.3 present the results of estimating models similar to those estimated in Tables 12-14, but with using instrumental variable regression instead of logistic regression. Each table includes both heteroscedasticity-consistent standard errors (top) and standard errors clustered at the firm level (bottom). In all models, the estimated coefficients for interaction terms are the same sign as those presented in the main results. With heteroscedasticity-consistent standard errors, significance is maintained throughout. Moreover, the heteroscedasticity-consistent standard errors are quite similar to the (unreported) classical standard errors, which provides some evidence that the model is not misspecified. (King & Roberts 2015).

The reported significance levels are based on the clustered standard errors. The significance of some of the interaction terms is not maintained under clustering. This fact, combined with the low likelihood (1-2%) that any particular patent is sold, suggests that the reduced form conditional logistic regression model for which results are reported in the main text provides a better fit for the data.

**Table A1: Instrumental variable regressions by CNW-reported patent effectiveness overall and for exclusion.**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Likelihood of sale (percentage points) | | | |
| | Product | | Process | |
| | (1) | (2) | (3) | (4) |
| Patent scope | 1.684* | 1.661* | 1.707* | 1.651* |
| | (0.137) | (0.140) | (0.136) | (0.138) |
| | (1.006) | (0.986) | (1.023) | (0.990) |
| CNW Overall | 0.315 | | 0.666** | |
| | (0.035) | | (0.039) | |
| | (0.283) | | (0.313) | |
| CNW Exclusion | | −0.004 | | −0.355 |
| | | (0.051) | | (0.047) |
| | | (0.314) | | (0.283) |
| Scope x CNW Overall | 1.706** | | 1.202 | |
| | (0.145) | | (0.127) | |
| | (0.859) | | (0.934) | |
| Scope x CNW Exclusion | | 2.745 | | 2.042 |
| | | (0.170) | | (0.159) |
| | | (2.720) | | (1.844) |
| Observations | 207,037 | 207,037 | 207,037 | 207,037 |
| $R^2$ | -0.038 | -0.082 | -0.023 | -0.055 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All covariates normalized as z-scores. All models run with application year by technology center fixed effects. Top standard errors are heteroskedasticity-robust. Bottom standard errors are clustered by firm. Significance levels based on clustered standard errors.

**Table A2: Instrumental variable regressions by CNW-reported patent effectiveness for cross-licensing and negotiation.**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Likelihood of sale (percentage points) | | | |
| | Product | | Process | |
| | (1) | (2) | (3) | (4) |
| Patent scope | 1.678* | 1.674* | 1.694* | 1.669* |
| | (0.136) | (0.136) | (0.136) | (0.136) |
| | (1.008) | (1.008) | (1.006) | (0.994) |
| CNW Cross-licensing | 0.332 | | 0.604** | |
| | (0.039) | | (0.041) | |
| | (0.257) | | (0.270) | |
| CNW Negotiation | | 0.320 | | 0.354** |
| | | (0.041) | | (0.037) |
| | | (0.255) | | (0.179) |
| Scope x CNW Cross-licensing | −0.004 | | −0.347 | |
| | (0.170) | | (0.155) | |
| | (0.639) | | (0.402) | |
| Scope x CNW Negotiation | | −0.034 | | −0.907 |
| | | (0.159) | | (0.154) |
| | | (0.511) | | (0.638) |
| Observations | 207,037 | 207,037 | 207,037 | 207,037 |
| $R^2$ | -0.014 | -0.014 | -0.015 | -0.022 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All covariates normalized as z-scores. All models run with application year by technology center fixed effects. Top standard errors are heteroskedasticity-robust. Bottom standard errors are clustered by firm. Significance levels based on clustered standard errors.

**Table A3: Instrumental variable regressions by transfer type and market characteristics, clustered standard errors.**

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Likelihood of transfer (percentage points) | | | | | | | | |
| | Sale | | | Exit | | | Spinoff | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Patent scope | 1.461** | 1.498** | 1.454** | 0.540 | 0.551 | 0.534 | 0.426*** | 0.434*** | 0.424*** |
| | (0.114) | (0.114) | (0.114) | (0.066) | (0.066) | (0.066) | (0.078) | (0.078) | (0.078) |
| | (0.694) | (0.701) | (0.689) | (0.679) | (0.685) | (0.674) | (0.142) | (0.143) | (0.142) |
| Market thickness | | 0.564*** | | | 0.120*** | | | 0.209*** | |
| | | (0.021) | | | (0.012) | | | (0.015) | |
| | | (0.065) | | | (0.046) | | | (0.025) | |
| Market congestion | | | −0.428** | | | −0.415* | | | −0.127*** |
| | | | (0.024) | | | (0.014) | | | (0.017) |
| | | | (0.210) | | | (0.223) | | | (0.034) |
| Scope x thickness | | 0.971* | | | 0.367 | | | 0.093 | |
| | | (0.101) | | | (0.058) | | | (0.069) | |
| | | (0.504) | | | (0.373) | | | (0.079) | |
| Scope x congestion | | | −0.663 | | | −0.536 | | | −0.138 |
| | | | (0.077) | | | (0.045) | | | (0.053) |
| | | | (0.535) | | | (0.579) | | | (0.085) |
| Observations | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 | 339,769 |
| $R^2$ | -0.011 | -0.014 | -0.012 | -0.001 | -0.004 | -0.004 | -0.001 | -0.001 | -0.001 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
All covariates normalized as z-scores. All models run with application year by technology center fixed effects. Top standard errors are heteroskedasticity-robust. Bottom standard errors are clustered by firm. Significance levels based on clustered standard errors.

## Appendix B: Data Construction

### B.1 Supplemental discussion of variable definitions and sample selection

Section B.1 provides additional details regarding the sample of patents analyzed in Section 2. The sample includes all patents filed 2004-2006 with four exceptions. Although these restrictions impose some limitations on the external validity of the findings in this thesis and although not all of these restrictions are strictly necessary, the following exclusions aid in interpretability and in making the strongest possible case for valid causal inference. First, the sample excludes patents that were not published prior to issuance. Patent applicants can elect not to publish a patent application prior to its issuance as a patent, although in practice applicants choose non-publication for only about 7% of patents. (Graham & Hegde, 2015). These patents are excluded from the sample due to the inability to observe their features at the time they were filed, which as discussed in Section 4.2 is important for the identification strategy. Second, the sample excludes patents examined within USPTO Technology Center 1600, responsible for examining biotechnology inventions, because the instrument has not been validated for use with biotechnology patents. (Kuhn and Thompson 2017). Third, the sample also excludes all patents that claim priority to a previously-filed patent application in order to avoid introducing endogeneity into the instrument. (Kuhn and Thompson 2017). Fourth, because this paper examines the effect of technology transfer between firms, the sample excludes any patent not already assigned to a firm at the time of issuance. The conclusions in this thesis are thus limited to original patents filed by firms in technologies other than biotechnology and published pre-grant.

The initial date of the sample was selected to capitalize on a change in U.S. patent law. Patents filed in or after 2001 are published by default after 18 months, allowing me to observe the claims of those patents not only at the time of issuance, but also at the time of filing. This in turn allows the construction of a measure of the causal instrument. Publicly available patent metadata is more complete for patents issued in 2005 or later, so the initial date of the sample is restricted to patents filed during or after the beginning of 2004.

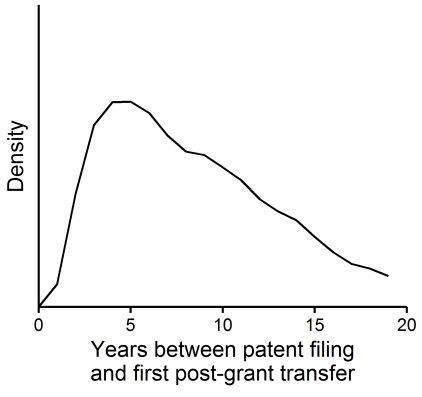



**Figure B1: Density of years between patent filing and first post-grant transfer.**

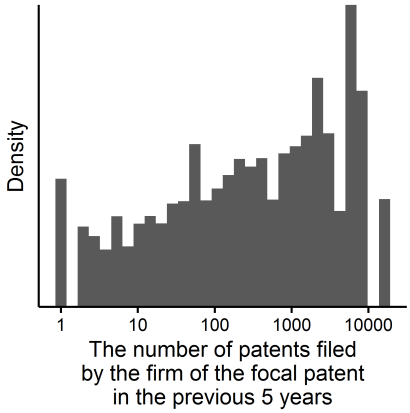



**Figure B2: Density of firm size on the patent level.**

The terminal date of the sample was selected to mitigate rightward truncation, allowing sufficient time for patents to issue and potentially be sold. Figure B1 shows the density of the lag between patent filing and the first post-grant patent transfers for patents filed in 2001 or earlier. Because the patent assignment data is complete through 2015, each patent in this sub-sample has at least 14 years to potentially be transferred, and most have even more time. Few patents are transferred so long after filing. Nevertheless, the long lag for some initial post-grant transfers illustrates that any analysis will be subject to at least some rightward truncation since data availability limits the identification strategy to patents filed in 2001 or later. In order to minimize rightward truncation as much as possible, the sample is limited to patents filed 2006 or earlier. Although the assignment data includes assignments filed in 2015, the final year of data is used for transfer-type classification purposes only. Thus, each patent in the sample is at hazard of being assigned up until the end of 2014, giving each patent at least 8 years to be transferred.

Firm size—variable 9 in Table 1—is operationalized as the log of the number of patents the firm received in the five years preceding the filing of the focal patent. Patent counts are used to measure firm size for several reasons. Although firm size can be measured in different ways, the relevant aspect of firm size for this analysis is not the firm's headcount or financial footprint but rather the extent of the firm's participation in the patent system. Also, systematic data measuring the size of small, private firms is generally lacking, so measuring a firm's size based on its footprint in the patent system provides consistency across observations. Firm size is calculated as of the filing year of each patent to capture changes over time. Figure B2 shows a histogram of firm size on the patent level.

The tenth, eleventh, twelfth, and thirteenth variables in Table 1 measure counts of different types of forward citations received by the focal patent. Patents in the sample receive on average 8.4 total citations and about .8 citations by the firm that owned the patent at the time of issuance. Total and self citations were identified from USPTO XML data. A blocking citation (102 or 103) indicates an instance where a patent examiner cited the focal patent to reject the claims of a later-filed patent. As described in Section 4.2, the scope of most patent applications is narrowed during examination at the patent office by a process of negotiation between the patent applicant and the patent examiner. Typically, the patent examiner rejects the patent's initial claims as being too broad, and the applicant responds by amending those claims to encompass a narrower invention. Blocking citations are the evidence the examiner provides to support such claim rejections. A claim is rejected under 35 U.S.C. 102 for failing the novelty requirement if the claimed invention is entirely described in a single prior art reference, while a rejection under 35 U.S.C. 103 is made when the claimed invention is obvious in view of some combination of prior art references. Blocking citations were identified by converting more than 50 million pages of USPTO correspondence from image to text and then using regular expressions to extract the patent numbers for the blocking citations from the resulting text files. Thompson and Kuhn (2016), who use blocking citations as a source of exogeneity to analyze the effects of winning a patent race, provide more detail on both the identification and significance of blocking citations.

The sixteenth and seventeenth variables in Table 1 describe characteristics of the local market for technology in which the patent is situated. A market definition based on a company-level indicator, such as an SIC code, would be too coarse to examine heterogeneity across markets for different types of technology because many large firms span technolog-

ical boundaries. Instead, I rely on internal classification of patents by the USPTO. When the USPTO initially processes a patent application, a reviewer assigns the application for examination to one of 531 art units based on the technology described and claimed in the application. The art units, which are split among eight Technology Centers, each include a number of patent examiners skilled in examining a particular type of technology, such as "Internal Combustion Engines" or "Wells and Earth Boring." I define each patent's local technology market as the USPTO art unit which issued the patent. For the analysis in Table 12, Table A3, and Table A4, I linked many of the art units to the industry categories identified by Cohen et al. (2000) ("CNW") to these USPTO art units by manual inspection. Some USPTO art units do not correspond to any category in CNW and are therefore are not linked. Similarly, some categories in CNW are either too broad (e.g., "2411: Basic Chemicals") or too specific (e.g., "3212: Precision Instruments" to be linked to a distinct USPTO art unit or group of art units. Patents in art units not linked to an industry category were omitted from regressions using CNW variables. The linkages, along with the number of observations and the percentage of patents transferred post-grant via sale, spinoff, or exit are shown in Table B2.

**Table B1: Correlation Matrix.**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sale (1) | 1.00 | | | | | | | | | | | |
| Exit (2) | 0.21 | 1.00 | | | | | | | | | | |
| Spinoff (3) | 0.10 | 0.14 | 1.00 | | | | | | | | | |
| Firm size (4) | -0.07 | -0.04 | -0.07 | 1.00 | | | | | | | | |
| Cites total (5) | 0.04 | 0.05 | 0.01 | 0.01 | 1.00 | | | | | | | |
| Cites self (6) | -0.02 | 0.00 | -0.01 | 0.06 | 0.45 | 1.00 | | | | | | |
| Cites 102 (7) | 0.03 | 0.03 | 0.00 | 0.03 | 0.15 | 0.07 | 1.00 | | | | | |
| Cites 103 (8) | 0.01 | 0.00 | 0.01 | 0.03 | 0.18 | 0.07 | 0.37 | 1.00 | | | | |
| Familiarity (9) | -0.01 | -0.01 | -0.01 | 0.00 | 0.04 | 0.08 | 0.02 | 0.01 | 1.00 | | | |
| Originality (10) | 0.01 | 0.02 | 0.00 | 0.01 | -0.03 | -0.03 | -0.01 | -0.01 | -0.36 | 1.00 | | |
| Thickness (11) | 0.05 | 0.02 | 0.03 | -0.16 | 0.05 | -0.02 | 0.00 | 0.02 | -0.05 | 0.00 | 1.00 | |
| Congestion (12) | -0.02 | -0.05 | -0.01 | 0.02 | -0.01 | 0.02 | 0.03 | 0.02 | 0.09 | 0.00 | -0.26 | 1.00 |

Table B1 presents the Pearson correlation between each variable. Due to the large sample size, nearly all correlations are statistically significant. The correlation matrix reveals several meaningful descriptive relationships. First, areas of higher congestion also have on average lower market thickness. Second, patents on technology that is more familiar to a firm are also less original on average when compared to patents outside the firm. Third, patents originally owned by larger firms are less likely to be transferred.

**Table B2: Summary statistics by CNW industry.**

| Industry | Obs. | Sale | Spinoff | Exit | Art Units |
|---|---|---|---|---|---|
| 1500:Food | 4,888 | 1.29% | 0.72% | 0.16% | 1791, 1792 |
| 1700:Textiles | 1,502 | 1.33% | 3.66% | 2.66% | 3765 |
| 2100:Paper | 742 | 31.27% | 0.54% | 30.46% | 1741, 3782 |
| 2200:Printing/Publishing | 1,761 | 2.16% | 1.76% | 0.62% | 3725 |
| 2320:Petroleum | 4,795 | 3.59% | 2.63% | 0.42% | 1797 |
| 2400:Chemicals, nec | 25,781 | 1.77% | 1.34% | 0.55% | 1700 |
| 2411:Basic Chemicals | 0 | | | | |
| 2413:Plastic Resins | 0 | | | | |
| 2423:Drugs | 0 | | | | |
| 2429:Miscellaneous Chemicals | 0 | | | | |
| 2500:Rubber/Plastic | 3,310 | 2.05% | 1.30% | 0.39% | 1711, 1712, 1732, 1734, 1742, 1743, 1747, 1761, 1762, 1763 |
| 2600:Mineral Products | 0 | | | | |
| 2610:Glass | 0 | | | | |
| 2695:Concrete, Cement, Lime | 256 | 0.39% | 1.95% | 0.00% | 1731 |
| 2700:Metal, nec | 4,579 | 2.64% | 0.39% | 0.98% | 1726, 1733, 1735, 1736, 1793 |
| 2710:Steel | 0 | | | | |
| 2800:Metal Products | 2,360 | 1.69% | 1.10% | 0.25% | 3726, 3729 |
| 2910:General Purpose Mach. | 0 | | | | |
| 2920:Special Purpose Mach. | 0 | | | | |
| 2922:Machine Tools | 4,484 | 1.27% | 0.83% | 0.33% | 3721, 3722, 3723, 3724 |
| 3010:Computers | 30,199 | 11.30% | 1.26% | 0.79% | 2100, 2400 |
| 3100:Electrical Equip. | 0 | | | | |
| 3110:Motor/Generator | 10,492 | 2.57% | 0.66% | 0.30% | 3745, 3746, 3747, 3748 |
| 3210:Electronic Components | 24,249 | 2.95% | 1.41% | 0.64% | 2827, 2831, 2832, 2833, 2834, 2835, 2836, 2837, 2838, 2842, 2843, 2844, 2845, 2847, 2848, 2859 |
| 3211:Semiconductors and Equip. | 17,741 | 3.46% | 0.83% | 0.32% | 2812, 2813, 2814, 1818, 2822, 2823, 2825, 2826, 2891, 2892, 2893, 2894, 2895 |
| 3220:Communications Equip. | 45,297 | 5.87% | 2.06% | 0.72% | 2600 |
| 3230:TV/Radio | 8,071 | 3.09% | 2.30% | 0.57% | 2422, 2486, 2612, 2622, 2651, 2656, 2661 |
| 3311:Medical Equip. | 7,458 | 4.16% | 2.15% | 3.35% | 3731, 3732, 3733, 3735, 3736, 3737, 3738, 3739, 3766, 3768, 3769, 3773, 3774, 3775, 3776, 3777, 3779 |
| 3312:Precision Instruments | 0 | | | | |
| 3314:Search/Navigational Equip. | 3,005 | 2.70% | 0.87% | 0.00% | 3646, 3655, 3661 |
| 3410:Car/Truck | 0 | | | | |
| 3430:Autoparts | 4,201 | 1.98% | 0.71% | 0.02% | 3611, 3616, 3618 |
| 3530:Aerospace | 1,866 | 1.66% | 1.23% | 2.52% | 3641, 3644 |
| 3600:Other Manufacturing | 0 | | | | |

### B.2 Patent reassignment data

Under U.S. law, every patent is by default owned as common property equally and indivisibly by its inventors. However, most inventors agree to assign their inventions to their employers as a provision of their employment agreement. Both this initial assignment and subsequent purchases of a patent are formalized through an assignment agreement. An assignment agreement is a simple legal document that lists the patent right that is transferred, the assignor (seller), and the assignee (buyer). The assignee can, but is not required, to record the transfer with the USPTO, which logs the details of the assignment in a publicly available database when and if the patent is published.

In practice, patent purchasers have strong incentives to record assignments since failing to do so creates a risk that they will lose title to their property. Patent assignments are legal instruments similar in some respects to deeds to real property and certain forms of personal property such as automobiles. Failing to record a transfer with a government agency creates ambiguity as to the legal owner of the asset. For example, suppose that an entity A transferred a patent to entity B, but that entity B failed to record the assignment. If entity A then transferred the patent to entity C, which did record the assignment, then entity C becomes the legal owner of the patent. Entity B could potentially recover the patent through litigation but would need to overcome the basic presumption of ownership created by the valid assignment entity C recorded.

I downloaded the complete database of assignment data from 1975-2015 from the USPTO. Each assignment includes a text field that describes the type of assignment. Following Serrano (2010), I first employed textual analysis to exclude from consideration assignments indicative of non-substantive transfers such as security agreements, firm renamings, and bequests. For example, I excluded assignments for which the text field included words such as "lien," "security," "mortgage," "satisfaction," or "reversion."

Next, I developed a novel disambiguation algorithm to create a unique identifier for each assignee and assignor. In the raw data, the parties are identified only by text name and address fields, both of which are subject to typographical errors or differences in how firms record their information. For instance, the same firm may record different addresses for different patents or may use an abbreviation for its name in one assignment but not in another. To address these imperfections in the data, I divided assignees by location—either U.S. state or a foreign country. I cleaned each name and address, correcting common misspellings and creating a canonical version of commonly abbreviated terms. I then grouped identical records together and gave each record within the same group the same firm identifier. Finally, I ran a clustering algorithm to coalesce groups where appropriate. Starting with groups having a small number of members, the clustering algorithm compared the focal group to all groups having a larger number of members. I used a combination of Levenshtein distance and Levenshtein ratio, depending on the length of the string, to compare both the name and address. If the match exceeded a threshold level, I coalesced the two groups by assigning the members of the smaller group the firm identifier of the larger group.

The precise criteria used to identify a match are complex and were constructed via many iterative passes of the algorithm in conjunction with manual review to identify the rates of Type I and Type II errors. However, both *ex ante* and *ex post* validation suggest that the algorithm identifies firms with a high degree of accuracy, with both types of error rates in

the low single digits.

Figure B3 shows the different types of patent transfers as a proportion of all transfers. Original transfers from the inventor to a firm comprise about 88% of all transfers since most patents are initially assigned to firms and not transferred again. Sales between ongoing companies make up the bulk of the secondary market, accounting for 6% of all transfers. Less common, but still sizeable, are transfers indicative of spinoffs and firm exits, accounting for 3.8% and .8% of all transfers. Finally, .5% of transfers are indicative of renaming and are omitted from subsequent analysis. The transfers shown in Figure B3 include those that occur both before and after patent grant. However, this thesis focuses on the relationship between patent scope and patent sales. Since a patent's scope is not determined until shortly before the patent is granted, the analysis in this thesis is limited to post-grant transfers.

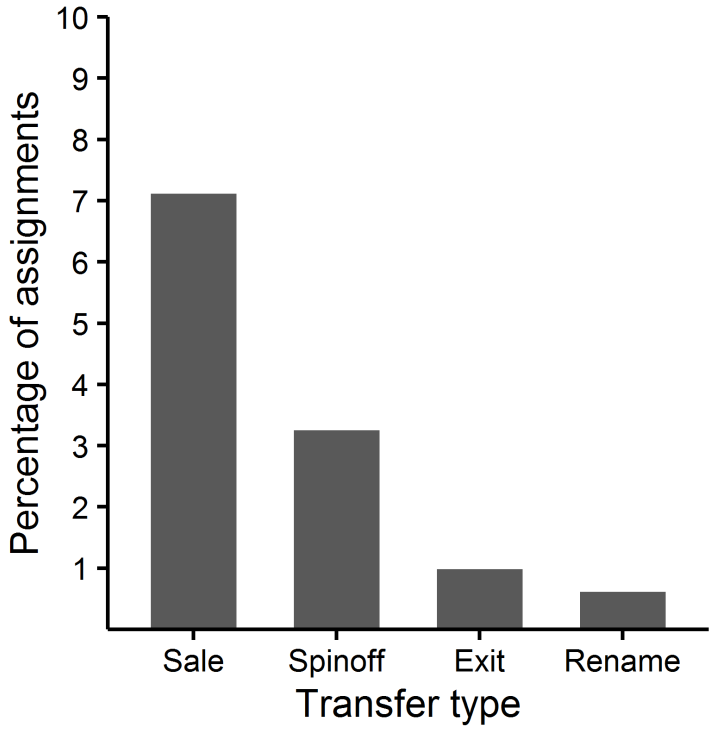



**Figure B3: Patent transfers by type as a percentage of all transfers, with initial transfers from inventors to firms omitted.**

### B.3 Patent congestion measure

The literature has not yet settled on an accepted technique for measuring patent rights congestion (i.e. patent thickets). In existing work, congestion-related concepts frequently turn on relationships between patent citations and patent ownership. For example, Cockburn et al. (2010) measure patent fragmentation based on the rate and incidence of citations made between patents owned by different firms. Von Graevenitz, Wagner, and Harhoff (2011) recently provided a novel technique for detecting the presence of patent thickets. They argued that the frequency of citation triangles—instances in which three firms all cite each other's patents—indicate the likelihood of overlap between patent rights in a particular technology.

However, Kuhn and Younge (2016) demonstrate that backward patent citations suffer from substantial selection effects and exhibit a secular time trend indicating near-exponential growth, raising questions of bias for most citation-based measures built using recent data.

I measured patent market congestion using patent-to-patent textual similarity data developed and validated by Younge and Kuhn (2015). In that project, Younge and Kuhn applied a vector space similarity model (VSM) based on term-frequency inverse-document frequency (tf-idf) weighting to calculate the textual similarity to every possible patent-to-patent comparison, involving over 14 trillion calculations and producing approximately 9 terabytes of data. Each calculation yields a cosine similarity value, normalized to a value between 0 and 100, that indicates the degree of textual similarity between the full technical description of pair of patents. The vector space model ("VSM") is one of the most common and robust models in the field of Information Retrieval Theory. (Manning, Raghavan, & Schtze). Originally developed for the SMART information retrieval system (Salton, Wong, & Yang, 1975), vector space models are used to construct a high-dimensional space based on the text of a given sample (or "corpus") of documents and now motivate the operation of nearly every modern-day search engine. The dimensionality of the vector space model was constructed from the originating vocabulary used in patent documents, not the assignment of patent classifications. Doing so allows the capture of similarity between patents whose patent classifications do not overlap, a key concern raised by Bloom, Schankerman, and Van Reenen (2013).

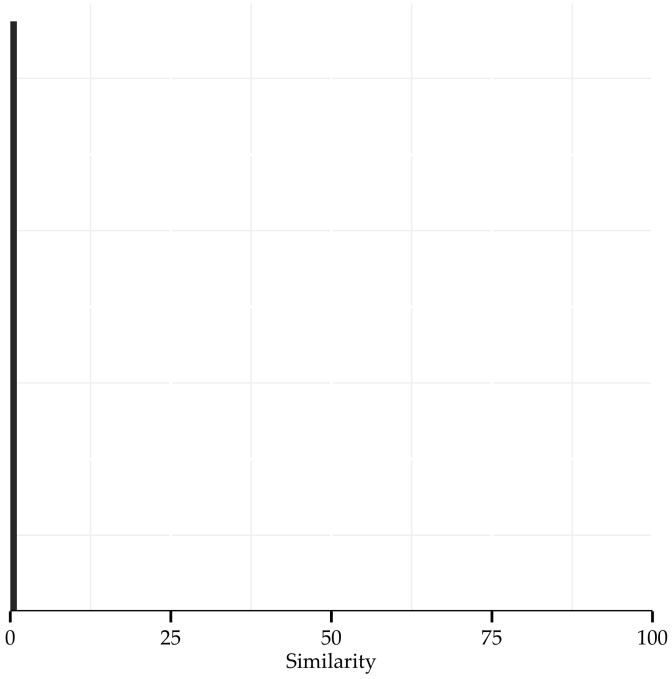



**Figure B4: Density plot of raw similarity values, from Younge and Kuhn (2015).**

Younge and Kuhn (2015) validate the patent-to-patent similarity measure using a range of approaches. The patent-to-patent similarity measure was an extremely significant predictor of

external ratings provided by a patent attorney, by a subject matter expert, and by the average response from a crowd of lay-person reviewers. The patent-to-patent similarity measure also predicted whether two patents share a patent class and subclass better than pre-existing techniques. Finally, relative patent-to-patent similarity values correspond intuitively with other constructs such as patent citations, patent families, and the degree of technological focus of a firm.

Since a randomly selected pair of patents is unlikely to have much in common, the resulting distribution is highly skewed. Figure B4, copied from Younge and Kuhn (2015), plots a histogram of the full population of 14.0 trillion values with a with a sharp spike at zero related to the 288.2 TB of calculated results that fell below the threshold of 0.10 and were therefore assumed to be zero, as well as the 8.1 TB portion of the distribution that fell above the threshold of 0.10 and was therefore retained. Figure B5 plots the distribution above 0.10 after the raw cosine similarity score was transformed by the arcsine of the square-root to improve interpretability.

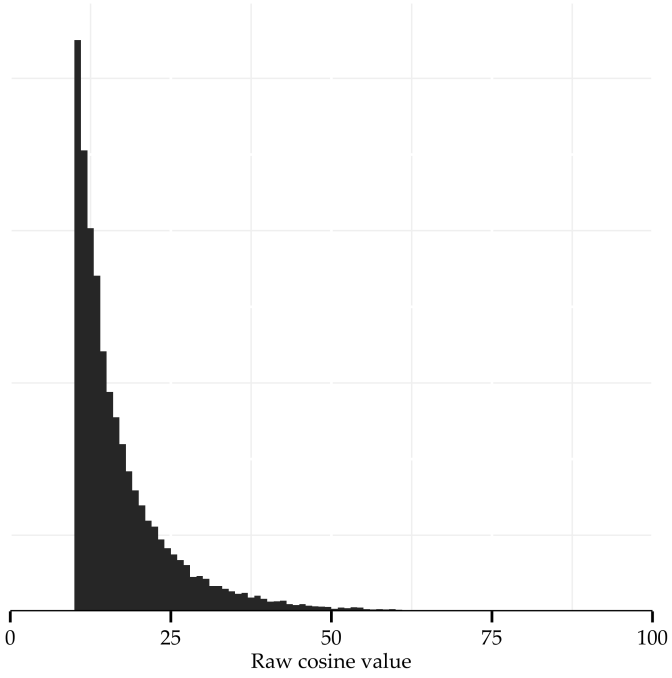



**Figure B5: Density plot of transformed raw similarity values above 10, from Younge and Kuhn (2015).**

The fourteenth and fifteenth variables in Table 1 were constructed based on this patent-to-patent textual similarity data. Familiarity measures the extent to which the focal patent is similar to past work by the same firm. Familiarity is operationalized by finding, for every patent, the maximum textual similarity to any other patent previously filed by the same firm. The resulting raw values are normalized to produce a z-score having mean zero and standard deviation one. Originality measures the extent to which the focal patent is different than past work by other firms. Originality is operationalized by finding, for every patent, the

maximum textual similarity to any other patent previously-filed by a different same firm. The resulting raw values are normalized and multiplied by negative one. Figure B6 plots the patent-level density of the maximum similarity to any earlier-filed patent, both within firm and between firms. As would be expected, same-firm similarity has a higher variance than between-firm similarity—firms sometimes branch out into technology new to the firm, but most patents tend to be somewhat similar to at least some prior work by other firms.

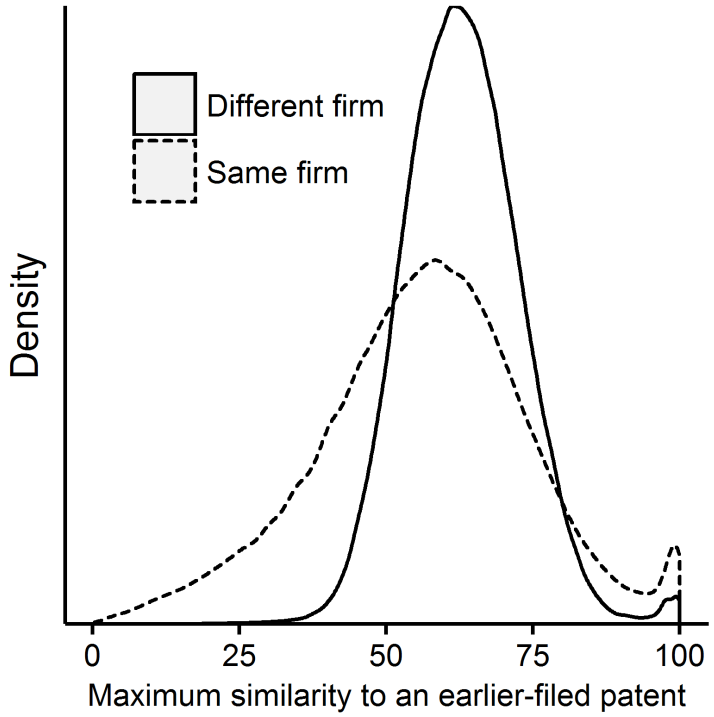



**Figure B6: Patent-level density of plot of the maximum similarity value to a previously-filed patent.**

I developed a measure of congestion based on the likelihood of overlapping patent rights between different firms in the spirit of Cockburn et al. (2010) and Von Graevenitz et al. (2011). For each patent in the sample, I first found the maximum patent similarity value between the focal patent and any other patent that is filed before the focal patent and is assigned to a different firm as the focal patent. This value proxies for the extent to which the focal patent is likely to conflict with a pre-existing patent right. I then determined the mean of these raw values for each art unit. Each focal patent received the raw value corresponding to its art unit, and the raw values were normalized across the entire distribution of patents to produce a z-score.

### B.4 Calculation of tf-idf

For clarity, this section explains in greater detail how the weighted vector was calculated for each patent based on the term-frequency, inverse-document-frequency (*tf-idf*). *Tf-idf* was calculated as: $tf - idf(t, d, D) = f_{t,d} * log(N/nt)$.

where $f_{t,d}$ = raw frequency of term $t$ in document $d$, $N$ = total number of documents, and $nt$ = the number of documents with term $t$. When combined into a single vector, the pattern of $tf - idf$ weighted values acts as a signature that differentiates (positions) one patent relative to another within the overall vocabulary space (Aizawa 2003). The *tf-idf*) weighting identifies particularly representative terms for a given patent, such that one would expect $tf - idf$ weighting to far outperform an ad hoc selection of keywords by an individual doing a keyword search. Any given patent will not use most terms in the vocabulary space; moreover, the formula for *tf-idf*) pushes the weight for a dimension to zero as a term appears in more documents (i.e., the ratio $N/nt$ approaches 1 and the logarithm of $N/nt$ goes to zero). Consequently, vectors are "sparse" (mostly zeros), or weighted toward zero. To illustrate this point, Figure B7 plots the cumulative percent of total *tf-idf*) weighting as a Lorenz curve for a random sample of 10,000 vectors. Vectors, on average, have 307 non-zero dimensional weightings, but the distribution of weightings is highly skewed; 20 dimensions contribute a majority of the weight going into the cosine calculation (Gini coefficient=0.705).

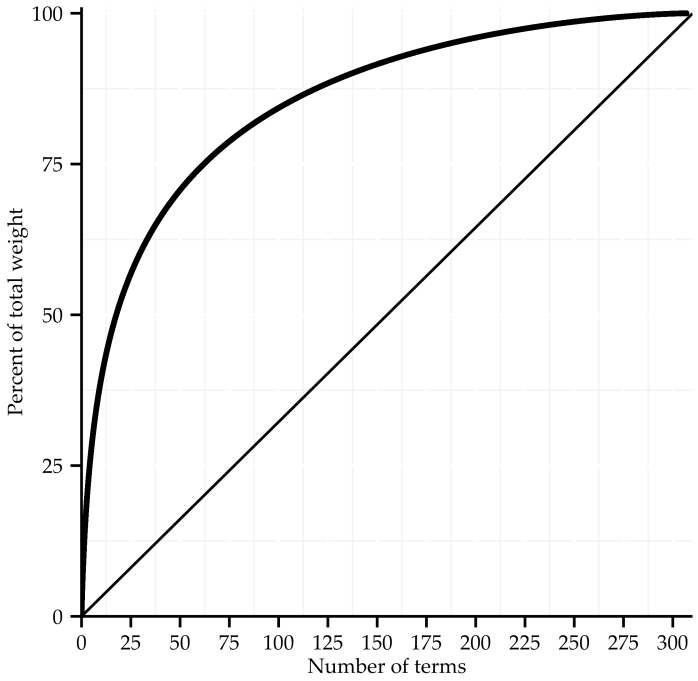



**Figure B7: Relative weight of *tf-idf*) terms in a random sample of patent vectors.**