The use and misuse of patent data: Issues for finance and beyond*

Josh Lerner
Harvard University

Amit Seru
Stanford University and Hoover Institution

This Draft: February 2021
First Draft: December 2013

Abstract

Patents and citations are powerful tools for understanding innovation increasingly used in financial economics (and management research more broadly). Biases may result, however, from the interactions between the truncation of patents and citations and the changing composition of inventors. When aggregated at the firm level, these patent and citation biases can survive popular adjustment methods and are correlated with firm characteristics. These issues can lead to problematic inferences. We provide an actionable checklist to avoid biased inferences and also suggest machine learning as a potential new way to address these problems.

*Both authors are affiliates of the National Bureau of Economic Research. We thank for helpful comments Jean Barrot, Shai Bernstein, Nick Bloom, Umit Gurun, Adam Jaffe, Andrew Karolyi, Bill Kerr, Adrien Matray, Scott Stern, Noah Stoffman, Per Stromberg, Xuan Tian, Heidi Williams, two anonymous referees, and participants in the American Economic Association annual meetings, and seminars at Cornell University, Harvard University, and the National Bureau of Economics Research. We especially thank Yuan Sun and Jinpu Yang for thoughtful analysis and outstanding research assistance. We also thank Paul Matsiras, Lilei Xu, and Yao Zeng for excellent research assistance with this paper; and Filippo Mezzanotti for the patent reassignment analysis. Harvard Business School's Division of Research (Lerner) and Center for Research in Security Prices at University of Chicago (Seru's previous affiliation) provided financial support. Josh Lerner has advised institutional investors in private equity funds, private equity groups, and governments designing policies relevant to private equity. All errors and omissions are our own. Send correspondence to Josh Lerner, 60 N Harvard St, Boston, MA 02163, USA. Telephone: (617) 495-6065. Email: jlerner@hbs.edu.

In the past several years, an increasing number of papers in the finance, accounting, and related literatures have made use of patent data. This growth has reflected the broadening of the topics seen as relevant to corporate finance researchers. Not only is innovation critical in many cases to firm survival—witness the fates of firms that failed to innovate successfully, such as Kodak, Motorola, and Xerox—but it illustrates the critical issues that motivate corporate finance theory more generally. Topics such as uncertainty, information asymmetries, and the intangibility of assets are central when it comes to financing innovative firms and projects.

The growth of interest in this topic among finance researchers can be seen, for instance, from a compilation of Google Scholar. We look at the number of papers each year that both cite at least one of the "top three" finance journals—the *Journal of Finance*, the *Journal of Financial Economics*, and the *Review of Financial Studies*—and contain the phrase "patent citation[s]" (the references in patents to earlier work added by patent examiners and inventors). We find that the share of such papers among those citing one of the top three finance journals rose from 0.1% in the 1990s to 0.5% in the 2000s to 1.7% between 2010 and 2019 (reaching 2.6% between 2018 and 2020). In the Appendix, we list over 80 papers using patent data, which have appeared in one of the top three journals between 2005 and 2020.

In many cases, the papers have used these data to shed fresh light on important problems. But in other instances, the interpretation of the results has been marred by a failure to understand some of the peculiarities of patents and patent data, which have led to conclusions that are not robust to the use of alternative methodologies. The presence of such systematic mistakes is understandable. The patent application and review process is extremely complex. The construction and features of the key database used for patent research—which originated at the National Bureau

of Economic Research in 1999 and has been updated to 2006—have not been as fully documented as would be desirable.

This paper is an attempt to rectify this omission. This paper consists of three primary parts. The second section starts with articulating on a particular but ubiquitous feature of patent-level data that can lead to problematic inferences, particularly in firm-level studies such as are commonplace in corporate finance research. This issue stems from the interaction of (a) the truncation of patent data and (b) the changing composition of inventors in a region or sector. We document the reasons for these effects and how they can affect researchers. We also highlight the two broad classes of corrections used to address these issues, and how they can be inadequate.

In the third section, we explore the consequences of truncation and a changing patent mix in patent and citation data for firm-level analyses. We introduce two methodological concepts. The difference in the actual patents granted relative to what was recorded in earlier data is what we call "patent bias." Similarly, we define "citation bias," the difference in citations to patents in earlier data with the citations garnered by the same patents over a longer period.

In an empirical illustration, we compare information on patent grants to publicly traded firms in the 2006 NBER patent database with the newer data on patents granted to the same firms applied for during the same time period. The newer, more complete data is collected through the end of 2012 using the method employed in Kogan et al. (2017). It therefore gives us a time window post-2006 to assess if patents that were applied for in earlier years were eventually granted. We first demonstrate that patent and citation biases are large and systematic: they are present more dramatically in recent years, in some technology classes, in some industries, and in some regions. We show that the popular methods in the literature to account for these biases only partially adjust for them, especially when it comes to citation data.

After characterizing these biases in detail, we explore how they impact inferences when analyzing patenting activity at the firm level. One solution for accounting for these biases at the firm level is to ignore them. The rationale could be that, when these biases are aggregated at the firm level, they end up being classical measurement error: i.e., they do not impact coefficients of the explanatory variables when patents or citations are used as dependent variables.

We show that this is unfortunately not the case. In particular, these biases at the firm level are strongly correlated with firm characteristics that are of key interest to researchers, both when we look at unadjusted and adjusted measures. In particular, firm size (market capitalization), the market-to-book ratio, the R&D-to-sales ratio, the ratio of cash to total assets, leverage, and return on assets are all positively associated with patent and citation bias, while the bid-ask spread is negatively associated with such biases. These biases are also positively related to the intensity of patenting in the given technology class. Thus, in many empirical settings where firm-level innovation is explored, several inferences about the phenomenon under study might be driven by non-classical measurement error. For instance, a one standard deviation increase in log firm size is associated with about a 0.02 standard deviation increase in adjusted patent bias and about a 0.10 standard deviation increase in citation bias. This issue affects virtually every study in the finance literature using relatively recent patent data.

In the final section, we grapple with how these concerns can be addressed in practice. We take both a short- and long-term perspective. We begin with a checklist that finance and other management researchers may wish to use as they formulate a research project using patent data. These items can help provide a robustness analysis to address the potential problems that we highlight.

We then turn to an alternative approach that could be harnessed to address patent and

citation bias: the use of machine learning (ML). We ask if exploiting a richer set of data at firm level than what patent-level adjustments use might help in reducing these biases. We present two sets of analyses. The first is targeted to address the patent bias at the firm level and the second one targets citation bias at the firm level. In each case, we seek to minimize the biases documented above. We assess performance by comparing the predicted values derived using activity through 2006 in the NBER data to the actual values as recorded as of the end of 2012, an approach similar to many of the adjustments discussed above.

In general, we find strong performance of the ML approaches. We show that adjustments generated by a variety of machine learning-based models using firm-level information–in addition to patent-specific information that is employed for popular "patent-based" adjustments in the literature–minimize patent and citation bias at the firm level. While the analysis is not the last word on this approach, it underscores the importance of using firm-level information when adjusting for patent and citation bias and points to the power of machine learning in addressing these issues.

It may be argued that the issues confronting users of patent data are similar to ones that those analyzing almost any contemporaneous database face. For instance, issues of truncation bias are commonplace: for a discussion of these issues in research into financial misconduct, see Dyck, Morse, and Zingales (2010) and Karpoff et al. (2014). But the dramatic changes in the direction and location of technological innovation (and patenting practice) over recent decades have led to a situation where these data limitations lead to biases in the results of patent-based analyses. Given the frequency with which these issues have surfaced in the published articles and working papers using patents in the finance literature, it is the goal of this article to document these biases' characteristics and consequences.

This paper overlaps to some extent with Dass, Nanda, and Xiao (2017). Our paper takes a more general approach to the issues discussed in that paper and provides a more detailed description of the biases and why they survive popular adjustments in the literature. We also put more emphasis on actionable steps to address these issues, as well as provide guidance to researchers seeking new ways of adjusting for firm-level patent biases.

It is also worth highlighting what this paper does not do. It is not a review of the key empirical features of patent grants and their economic applications: Jaffe and Trajtenberg's classic volume (2002) remains the "go to" reference for such an analysis (also see the brief review in Online Appendix A). Nor is it a review paper summarizing the crucial works using patent data. Far more details about the patent application process can be found in many legal texts. The paper does not attempt to explore the issues associated with non-U.S. patent data, whose use in some cases can alleviate some (though not all) of the issues highlighted here. This decision is rooted in the twin desires to keep this paper manageable in length and to address the fact that the overwhelming majority of the papers in the finance literature listed in Appendix 1 have analyzed U.S. data.

1. The Nature of Patent Data

Patent data are exceedingly informative. While firms need only report R&D in corporate filings if the expenditures are "material," all granted patents are recorded. Moreover, R&D expenditures are typically reported in aggregate, not broken down by product line or geography as patents are. Finally, R&D expenditures are an innovative input, rather than an output: the effectiveness of the research may vary tremendously.

The first, most fundamental U.S. database is from the U.S. Patent and Trademark Office (USPTO) itself. The database covers patents awarded between 1976 and today, with earlier patents only in PDF format. These data pose several issues. First, there is no identifier that uniquely flags each applicant. Moreover, a huge number of variant names of frequent patentees appear. So these data can be difficult to use.

The NBER Patent Citation Dataset—created under the leadership of Bronwyn Hall, Adam Jaffe, and Manuel Trajtenberg (HJT)—was designed to address these difficulties. The original database sought to capture the key information on each utility patent awarded between 1963 and 1999 in a readily accessible database. (About 90% of all patents issued are utility patents.)

The primary contribution of the NBER database was in its handling of patent assignees. In particular, the authors linked the first assignee of each patent (other assignees were ignored) to its Compustat CUSIP identifier, as long as it was a U.S. publicly traded entity. They used the 1989 Compustat file to do this, so the coverage deteriorates over time: for patents granted in the mid-1980s, about 65% of all patents with a U.S. inventor were matched, but among those granted in 1999, the share falls below 50%. This attrition reflects the entry into patenting of numerous firms that were not publicly listed in the 1980s. The authors also assigned patents owned by major operating subsidiaries to their patents using the 1989 edition of the *Who Owns Whom* directory. The main data set also tabulated the numbers of citations made and received (between 1975 and 1999).

There have been a number of updated versions of the NBER data, the most recent of which is the file extending through 2006, compiled under the leadership of Bronwyn Hall and Jim Bessen (Bessen (2009)). Among the changes were the inclusion of patents and citations through the end of 2006, all assignees to each award, and patents other than utility awards. The most significant

progress, however, was made on matching assignees to Compustat identifiers (using GVKEYs, the more "permanent" of the two firm identifiers used in Compustat). They extended the number of matches between the assignee names and Compustat by using a computerized algorithm that stripped suffixes and accepted some inexact but highly probable matches. They also sought to document when the assignee firms were subsequently acquired, and the associated GVKEYs of the acquiring entities. The 2006 data update did not, however, revisit the mapping between parent and subsidiary firms undertaken by HJT using the 1989 data.

Since the completion of the 2006 NBER database, there have been a number of efforts to update and enhance these data, including efforts to rationalize the names of individual inventors (Li et al. (2014)), update links to public assignees (Bena et al. (2017)), and analyze earlier patents (Moser and Voena (2012); Kogan et al. (2017)).[1]

2. The Central Challenge

2.A Truncation

The problems highlighted in the introduction may have seem like abstract ones, more of theoretical interest to economists of innovation than finance researchers. But the failure to control for the changing composition of patenting firms can cloud inferences about the impact of financial policies.

[1] The state of development of data from other patent offices is much less mature. The development of an EPO research database remains a work in progress: an initial mapping of UK firms' filings has been undertaken by Grid Thoma and co-authors (2010), as well as a mapping between the names in Bureau van Dyck's Amadeus dataset and European patent assignees (http://www.epip.eu/datacentre.php). While there have been recent efforts to make Chinese and Japanese data available online as well, this information remains much less well scrutinized.

We focus on a single but wide-spread issue. As we indicated in the introduction, the interaction between the truncation of patent awards and the differences in patenting across regions and sector can pose difficulties. As we will discuss, while these problems are known to some researchers—and there are some popular ways to deal with these—it is difficult to account for these problems entirely when conducting firm-level analysis in corporate finance and related research. Knowing the nature of these biases, however, does allow one to ex ante predict how inferences in various empirical settings might be impacted.

These issues stem from two features of patent data. The first critical effect has to do with its truncation. The patent literature has generally focused on analyzing patent filings by the application year, rather than the award year. The motivation is that firms, eager to protect their intellectual property, will tend to file for patents soon after the discoveries are made. The gap between the date at which the patent is applied for and issued, however, is a product of many other considerations, such as the area of technology covered by the patent and the contemporaneous state of the patent office. To eliminate this noise, looking at patent by application date seems a more reasonable approach.

This adjustment, however, is not sufficient to account for the truncation problem. In particular, any analysis of patent filings near the end of the database needs to control for truncation. This is illustrated in Figure 1(b), which depicts the number of patents in the NBER 2006 dataset by application year. Because this database only reports the number of patents that issued by the end of 2006, there is a dramatic tail-off: the number of applications peaks in 2001. This has nothing

to do with the actual number of filings (which, as Figure 1(c) reveals, actually continued to rise steadily), but instead with the delays in issuing patents.[2]

This truncation issue is even more severe when it comes to computing citations. It is rare for a patent to be cited by another patent filing before the cited patent has issued. Even after issue, the reaction is not instantaneous, as the citing patents themselves have to work their way through the application process. As Figure 1(d) reveals (again drawn from the 2006 patent data), the rate of citations per patent peaked for patents filed in 1986 and began a rapid slide by the mid-1990s. As a result, more recent cohorts of patents will be mechanically less cited, even if their degree of innovativeness does not decline.

The truncation is not uniformly distributed across technology classes in which patents are generated. Figure 2(b) provides an illustration of the truncation issue, by reporting citations per patent for cohorts of electronics and chemical patents. In each case, there is a dramatic tail-off in citations in later years for the younger cohorts: for many of the patents in the younger cohorts, there has not been sufficient time for these patents to garner citations in the later years. We can anticipate that if we were to revisit this distribution for the younger cohorts in a later year, the distribution of citations by year will more closely resemble that of the older cohorts. But the tail-off is more dramatic for patents in the electronics subcategory.

There are two primary responses to the truncation issue seen in the literature. The first of these we term the "fixed-effect" approach. We term the simplest approach, pioneered by Jaffe and

[2] It might be argued that it would be even more defensible to look at the original patent filing date: that is, the date the original patent filing was made, before taking into account divisions, continuations-in-part, and the like. But given that the various versions of NBER patent databases do not readily allow such a determination, and that such a step would doubtless intensify the truncation problems that we discuss in this essay, such an alternative approach appears impractical.

Trajtenberg (2002) and HJT (2001, 2005), time adjusted: one estimates a distribution function of the patents and citations over time using non-truncated data, then infers what the truncated data should look like.

The time fixed effect adjustment relies on "re-scaling" firms' patent information with data about patent population during a certain period. In the adjustment for patents, the annual heterogeneity is removed by dividing the number of granted patent applications assigned to each firm in a year by the total number of granted patents applied for in a corresponding year:

$$Adj\ Patent_{ft} = \frac{n_{ft}}{N_t} \tag{1}$$

where $Adj\ Patent_{ft}$ is the adjusted number of granted patents applied for by firm f in year t, n_{ft} is the total number of granted patents applied for by firm f in year t, and N_t is the total number of granted patents applied for in year t.

Similarly, when adjusting citations, the annual heterogeneous component is corrected by dividing the number of citations received by each firm by the average number of citations received by patent cohorts in the same year:

$$Adj\ Citation_{ft} = \frac{\sum_i^{n_{ft}} Citation_i}{\sum_j^{N_t} Citation_j / N_t} \tag{2}$$

where $Adj\ Citation_{ft}$ is the adjusted number of citations received by firm f to granted patents applied for in year t, $Citation_i$is the number of citations received by ith patent from firm f, n_{ft} is the total number of granted patents applied for by firm f in year t. $Citation_j$ is the number of citations received by jth patent applied for in year t, and N_t is the total number of granted patents applied for in year t.

This approach was originally developed for the use with aggregate patent data, but has since been applied to individual patent data as well (e.g., Chemmanuer and Xuan's 2018 analysis

of hostile takeovers). It might be thought that such inferences would be extremely noisy at the individual patent level, as very small differences in early citation rates—especially if they vary across technology class and spatially—could be amplified through such an imputation approach. As suggested by Figure 2(b), differences across technologies over time may also pose problems. Finally, work by Nicholas (2008) and Kolev (2013) suggests that more fundamental patents have a much longer "half-life" of citations than more routine extensions, which might lead us to worry that such inferences could introduce systematic biases.[3]

A variant, also pioneered by HJT (2001), which we term time and tech class adjusted, is to look at patents and citations relative to those awarded in the same technology class and year. In this adjustment, the information about different patent classes in different years is also used in the adjustments. The number of patents in different class assigned for each firm in a year is adjusted with total number of patents applied in corresponding year and class:

$$Adj\ Patent_{ft} = \sum_{k}^{M} \frac{n_{fkt}}{N_{kt}} \quad (3)$$

where $Adj\ Patent_{ft}$ is the adjusted number of granted patents for firm f applied in year t, n_{fkt} is the total number of granted patents for firm f applied in year t in class k, and N_{kt} is the total number of granted patents applied in year t in class k. M is the total number of patent classes in the data. In our analysis, M equals 6, based on the HJT classification.

[3] A related approach is to only use citations in a short window after a patent award. For instance, Lerner, Sorensen and Stromberg (2011) only look at citations in the three years after awards. This avoids some of the issues delineated above, but early citations only capture a very small number of total citations. Thus, the information that is discarded through such an approach is potentially quite significant.

The same is true for citation adjustment: we divide the number of citations by patents in different classes for each firm in a year by average number of citations for patent cohorts in each patent class in the same year.

$$Adj\ Citation_{ft} = \sum_{k}^{M} \frac{\sum_{i}^{n_{fkt}} Citation_i}{\sum_{j}^{N_{kt}} Citation_j / N_{kt}} \quad (4)$$

where $Adj\ Citation_{ft}$ is the adjusted number of citations received by firm f to granted patents applied for in year t, $Citation_i$is the number of citations received by i[th] patent from firm f, n_{fkt} is the total number of patents for firm f applied for in year t in class k, $Citation_j$ is the number of citations received by j[th] patent applied for in year t, and N_{kt} is the total number of granted patents applied for in year t in class k. M is the total number of patent classes in the data. Again, M is 6, based on the HJT classification.

This methodology has also had widespread use in the finance literature. For instance, Seru (2014), in his analysis of the impact of conglomerates on innovation, uses the ratio of the number of citations per patent for each firm to the mean citations per patent in the same cohorts as the firm's patents. Ideally, this approach will control not just for truncation problems, but also adjust for the shifts engendered by changes in patent office policy and technological fluctuations. However, the approach is still subject to distortions: a single early citation may lead to a large ratio. The issue of important patents having differing citation profiles over time (as discussed above) is also a problem here. Moreover, as Hall and co-authors (2001) suggest, by undertaking such a normalization, one may be sweeping away information: for instance, if a key innovation leads to a substantial burst in innovation in a given industry.

A second class of approach is what we term the "quasi-structural" one. In summary, this method "models" the distribution of citations based on citing year effects, cited year effects, and

propensity to cite fixed effects for different technology classes (Detailed descriptions of this adjustment are in Online Appendix B.) For brevity, we produce analyses with adjustments using citing year and cited year effects. The analysis based on propensity to cite adjustment is relegated to Online Appendix B and the tables and figures there. This method also has its share of potential problems, since it is hard to model distributions over time, which may be altered as new patents and technologies are added in the models.

In conclusion, the time lag between the filing of a patent application and its subsequent grant results in a mechanical tail-off in patent grants towards the end of the sample. Moreover, it may be a decade or longer after a patent is filed before one can get a good sense of how influential it is from citations. While it is possible to adjust the number of patent grants and number of patent citations received in early years based on historical patterns—and thus project the total number of patents or amount of citations likely to be ultimately received—these estimates can be quite imprecise and potentially biased. To understand why, we need to explore the changing composition of inventors in a region or sector.

2.B Changing Inventor Composition

Were the composition of patents the same over time, the implications for corporate finance researchers from the truncation issues discussed in Section 2.A might be thought to be relatively modest. Some simple corrections could address these truncations. Rendering this problem particularly difficult, though, is the dramatic changes in the composition of inventors across regions and sectors seen over the past several decades. The interaction between truncation and changing inventor composition can lead to biases in firm-level analyses, as we discuss in Section 3. We explore the second of these factors in this section.

First, it is important to highlight that the past three decades have seen a dramatic acceleration in patenting activity in the United States and elsewhere in the world. Figure 1(a) depicts the number of patent awards in the U.S. in the NBER 2006 database and highlights the three-fold increase between 1975 and 2006. If we look at patent applications (whether ultimately successful or not) during the same period, there is a four-fold increase, as Figure 1(c) reveals. (This data series, unlike the others in the paper, is drawn from the annual reports of the USPTO, not the NBER database.)

Practitioner accounts suggest that this increase in patent filings was a response to the increase in patent rights, rather than a reflection of an endogenous shift in the amount of innovation. This shift towards a more "pro-patent" policy has been effected most dramatically through the decisions of the Court of Appeals for the Federal Circuit (Merges (1992)).

Were this a case where the increase was uniform, it again be addressable through a series of correction factors. But this secular trend towards more patenting is not uniform across technology classes or regions. Thus, simple adjustments such as time fixed effects will fail to fully account for such interactions, which can then bias inferences in predictable ways. In particular, the propensity to patent across technologies and industries varies dramatically in time-varying ways, and as a result the "density" of patents in given areas may be very different. Thus, some patents will be heavily cited due to their technological location, rather than their fundamental innovativeness.

These issues are illustrated in Figure 2(a), which compares awards assigned to the HJT "computers and communications" (henceforth computers) and chemicals classifications. The figure makes clear that computers experienced a much more dramatic run-up in patenting activity in the 1980s, 1990s, and early 2000s.

Figure 2(b) illustrates the differing truncation of patent citations across industries. It shows that the distribution of patent citations to computer firms is skewed leftward: more citations happen sooner after the award. The more rapid tail-off in the citations may be due to two factors: (a) the more rapid obsolescence of these technologies, leading to a reduced propensity to cite older awards, and (b) the more recent vintage of the typical patent in this area.[4] As we will demonstrate, these measurement problems can have very substantial implications for empirical analyses, especially ones that explore innovation across firms in similar industries using different technologies.

Relatedly, the rate of patenting has grown at different rates across regions. Many patent-based analyses exploit regional differences in order to identify effects: e.g., using the staggered adoption of policies by different states. Any analysis that hopes to explain differences in innovativeness across firms but does not control differential patenting across regions is likely to result in problematic inferences.

To illustrate the severity of the problem, we proceed here by comparing patents by assignees in a state that has frequently been on the cusp of business-friendly policy reforms, Delaware, with California and Massachusetts, which have been accused of being at the other extreme. Figure 3(a) reveals that patenting only increased between 1990 and 2000 by a few percent for inventors in Delaware, while for inventors in the other two states, the increase was two-and-a-half fold. Figure 3(b) shows that patents with assignees in California and Massachusetts were more

[4] Another contributing factor are differential lags between application and grant dates. Online Appendix C (Panel A of Table C1) shows the distribution of the lag (in years) between application and grant date for patent applications across technology classes. There is some heterogeneity across classes, with computer patents having the longest lag. Panel B of Table C1 in the Online Appendix shows the lag (in years) between citing and cited patents across technology classes. Again, there is considerable heterogeneity across technology classes.

likely to be cited. The figure also highlights that the drop-off in citations is more concentrated for the California and Massachusetts patents in the very last years of the sample.

Of course, behind these differences are considerable disparities in the industry composition of the firms active in these states. Computer and electronics firms are far more likely to be located in California and Massachusetts. Moreover, the mixture of industries across states changes over time. As we will demonstrate, not accounting for industry composition across states might lead us to spuriously conclude that these policies affected innovative activity at firm level.

3. The Consequences for Firm-Level Analyses

We explore the consequences of truncation and a changing patent mix in patent and citation data for firm-level analyses such as commonly seen in corporate finance. While several methods are available to account for biases due to time, technology class, and region, as discussed earlier, these methods are primarily for adjusting for biases at the patent level. Most research in corporate finance is at the firm level. We now demonstrate that the popular methods available to account for these biases at the patent level may not be sufficient when one aggregates patents at the firm level. In particular, we will demonstrate that the residual measurement problems that emerge are not pure noise, but rather are related systematically to firm characteristics. This makes it difficult to disentangle firm-level factors that truly impact innovative activity from spurious measurement error induced due to the problems discussed above.

To illustrate the nature of the problem, we estimate the firm-level bias that is created by truncation issues. We do so by computing the difference—both unadjusted and adjusted for truncation using popular methods—between (a) the patenting and citation activity of a firm in a given year as recorded by the end of the 2006 NBER data relative to (b) the patenting and citation

activity of the same firm in the same year as recorded in *"our data"* that tracks patents granted through the end of 2012. Our dataset is constructed by scraping the patent records directly from 1976 through 2012, using a similar procedure as in Kogan et al. (2017).

More specifically, we construct the unadjusted "patent bias" for each firm-year by comparing the number of patents for each firm filed in each application year in our data (thus, which have been granted by 2012) and in the NBER 2006 dataset (i.e., granted by 2006). It should be noted that this measure will understate the true extent of the truncation problem since there will be patents granted subsequent to 2012 based on applications from 2006 and before. We repeat a similar exercise to compute "citation bias" for each firm-year: we compare the number of citations to all the patents that each firm filed in each application year in our data (i.e., citations in patents granted by 2012 to applications filed by a firm in a given year and granted by 2006) and in the NBER 2006 dataset (i.e., citations in patents granted by 2006 to applications filed by a firm in a given year and granted by 2006). Finally, we also construct various versions of firm-level adjusted patent and citation bias measures. We do so by using methods that adjust for biases at the patent level, as discussed in the previous section.

Next, we undertake regression analysis to relate these firm-level biases to firm characteristics using a sample of publicly listed firms. We explore how these biases relate to the following characteristics: Firm Size (Log Size), Market Value to Book Value (Log M/B), R&D Investment to Sales ratio (Log RD/Sales), Cash to Assets ratio (Log Cash/Assets), Return on Assets (ROA), Market Leverage (Log Leverage), and Bid-Ask Spread (Log Spread). In addition, we explore how the firm-level biases might relate to geographic and technology class characteristics of the firm by accounting for the number of granted patents in the same class as the modal technological class of the firm's patents (Log (Patents in Technology Class)) and the

number of granted patents in the same state as the modal state of the assignees in the firm's patents (Log (Patents in State)). Online Appendix D details the exact definitions of these variables and how they were constructed. There are a total of 1807 publicly listed patenting firms in our sample, with 1443 firms having no missing information.

We analyze both unadjusted biases and adjusted ones. In the figures, we focus on all the popular methods used in the literature discussed in Section 2.A. Tables 1 and 2, in the interests of space, focus on just one adjustment method (time and technology). The results using the other methodologies (in Online Appendix E) look qualitatively very similar to those reported in the paper.

3.A Unadjusted and Adjusted Biases in Publicly Traded Firms across Time: Graphical Analysis

Figure 4A illustrates the firm patent bias over time. We sum the patent bias across publicly traded firms by application year. The resulting patent bias is more severe for more recent patents than older ones. This is because many patents that are applied for close to 2006 end up being granted between 2007 and 2012. While the adjustments—the time fixed effects and the time and tech class fixed effects—help alleviate some of the truncation problem, a significant portion of the bias remains. Several thousand missing patents remain unaccounted for, even after the adjustments. We also present the patent bias on a per firm basis, and find similar patterns.

Figure 4B shows the citation bias over time. We again sum citation bias across publicly traded firms by year, as well as taking the average per firm. Here, the trend is a bit different. The bias is most severe during the year 1998, which is eight years before the end of the sample period in 2006. Using information on citations granted to these patents for another six years past 2006—which implies that we track citations for these patents 14 years after issuance—yields a large

number of subsequent citations to these patents that were not captured until 2006. In contrast, the bias is not as large for patents granted as of 2006. This is likely because tracking citations for six years after issuance—that is between 2006 and 2012—is not long enough of a time period to capture the bulk of subsequent citations, as citations tend to peak with some lag (see HJT (2001)). Thus, we are likely severely underestimating the true extent of citation bias for the patents that are granted towards the end of the sample. While the adjustments are useful in alleviating some of the bias, a significant bias remains. In essence, adjustments using historical data do not fully account for the time-varying dynamics at the firm level in both patents and citations.

3.B Unadjusted and Adjusted Biases in Publicly Traded Firms across Technology Class: Graphical Analysis

Figure 5 (Panels A to C) shows the patent bias at the firm level for different technology classes, again unadjusted and adjusted. Firms are assigned to a particular technology class in a given year, based on the modal primary patent class of patents produced by the firm in that year (using the U.S. patent classification system). We then sum the bias across publicly traded firms in each technology class, as well as taking an average per firm. It is clear that the most severe patent biases are concentrated in the computer and the electronics classes. This reflects the explosion of patents in these sectors relative to other classes, especially towards the end of the sample.

Similar patterns emerge when we adjust the bias in Panel B and Panel C. Interestingly, when we adjust by the fixed-effect method, some classes display a "negative bias": the adjusted number of patents exceeds the actual number issued through 2012. This pattern may reflect the failure of the fixed effects to fully capture the rapid acceleration of computer-based patenting and the declining share of other classes.

Figure 6 shows the citation bias at the firm level for different technology classes. As before, we assign publicly traded firms into technology classes in each year and sum and average the citation biases. Compared with the large unadjusted citation biases in Panel A, the adjustments in Panels B through D do help. However, as was the case before, a significant part of the bias remains, particularly when it comes to computer patents. Finally, the pattern of citation bias peaking earlier in time than patent bias, discussed in Section 3.A, emerges across technology classes.

To the extent that citation and patent bias across technology classes illustrated in this section varies within and across firms, granted patents and their citations within and across firms will be less comparable as we get closer to the end of the sample period in the NBER dataset.[5]

3.C Unadjusted and Adjusted Biases in Publicly Traded Firms across Regions: Graphical Analysis

Figure 7 illustrates the distribution of patent bias at the firm level across different states. We assign firms to different states based on the modal US state or territory of the assignees recorded by USPTO at the time of the application.[6] We then sum the bias by state across publicly traded firms, as well as taking an average per firm. As can be observed, total patent bias is mainly concentrated in states like California, New York, Texas, and Washington. This reflects not only

[5] Another way to illustrate this issue is based on Online Appendix F. The figure shows the distribution of granted patent applications (Panel A) and the mean number of citations per patent (Panel B) for the six HJT technology classes. This analysis is undertaken at the firm level, with firms assigned to the modal class of patents granted to a given firm in a given year. The figure highlights, when aggregated at firm level, the substantial heterogeneity in the volume and time trends of patents and citations.

[6] In a few rare instances, firms report multiple assignee states for a patent. We randomly picked a state in such situations. Doing this procedure several times assured us that the inferences made in this section are not sensitive to this choice.

the size of the states (there are many more patent applications), but also the concentration of computer firms in these states. The patterns in the unadjusted data (Panel A) remain even after adjustment in Panels B and C. Again, we see "negative bias" in some states.

Figure 8 shows the distribution of citation bias at the firm level across different states. We again sum the citation bias in a state across publicly traded firms, with firms assigned to states each year as discussed above. Similar to patent bias, states like California, New York, and Texas suffer most from citation bias. Adjustments help to some degree, but significant bias remains. As we noted before, patent and citation bias by state is particularly evident among recent patents.

When we look at the average bias per firm on the state level, we find that, unlike in Figures 4 through 6, the calculations in Figures 7 and 8 lead to the identification of a different subclass of patents as those with the greatest absolute bias. In particular, we find that two states frequently are identified the largest absolute patent and citation bias: Idaho and Oklahoma. In both cases, the states are characterized by relatively few publicly traded firms. Moreover, they have very skewed patenting activity. A single firm, Micron Technologies, represented 80% of Idaho's patenting in many years over recent decades, and alone exceeded the patenting output of a number of states. Oklahoma is similarly characterized by a small number of oil-and-gas firms that are prolific patentees.

3.D Unadjusted and Adjusted Biases in Publicly Traded Firms across Industries: Graphical Analysis

As we have seen from our analysis so far, these biases occur more in some technology classes and regions than others. To the extent that firms in some industries are more active in certain technology classes and regions, we might expect these biases at the industry level as well.

In Online Appendix G, we plot the patent bias at the firm level across different industries, defined using the NAICS code and Standard Industrial Classification (SIC) codes. We assign firms into different industries at time of the patent application using Compustat. We then sum the bias by industry across publicly traded firms. Patent bias is mainly present in industries like manufacturing (especially class 33, which includes computer and communications equipment manufacturing), information technology, and services (which includes computer systems design and R&D services). The adjustments reduce the biases, but considerable distortions still exist. A comparison of Panels B and C shows that adding controls for technology class has very little impact on the biases across these industries: either these classes are too crude, or the differences across industries in patenting growth are largely orthogonal to the controls. A second analysis in Online Appendix G illustrates the citation bias at the firm level across different industries. The unadjusted citation bias is mainly concentrated in manufacturing, information technology, and services. It proves difficult to eradicate even with fixed-effect and quasi-structural adjustments.

3.E Unadjusted and Adjusted Bias and Firm Characteristics in Publicly Traded Firms: Regression Analysis

Our analysis so far has illustrated the bias in patent and citation counts at the firm level that varies across time, technology class, region, and industry. Moreover, popular patent-level methods for adjustment seem to be insufficient in eliminating this bias. At the firm level, another adjustment for such "measurement error" in patenting and citations could be to estimate regressions that account for time- and firm-invariant characteristics using time and firm fixed effects. We now show that this approach is not sufficient either. Consistent with our discussion earlier in this

section, this analysis will reveal that these firm-level biases vary in complex fashion across time, technology class, regions, and industries.

To illustrate this more formally, we estimate fixed-effect OLS regressions in Table 1. The unit of observation in each case is firm-year observations of patent bias between 1976 and 2006, with the results reported in six columns. The six columns employ as our dependent variables unadjusted patent bias (columns 1-3) and patent bias adjusted for time and technology class fixed effects (columns 4-6). While we pick the most stringent patent-level adjustment for presentation, our results are similar when we use other adjustments instead. The dependent variables are computed as the adjusted or unadjusted difference in log of one plus number of successful patents filed by a firm in a given year as of 2012 ("our data") and log of one plus number of successful patents filed by that firm in the same year as of the end of sample in the NBER 2006 dataset. Logarithms are taken to account for skewness in patenting activity.[7]

In Table 2, we employ a similar specification using unadjusted and adjusted citations bias as the dependent variables. In particular, the six columns employ as dependent variables unadjusted citation bias (columns 1-3) and citation bias adjusted for time and technology class fixed effects (columns 4-6). As before, while we pick the most stringent citation adjustment for presentation, our results are similar when we use other adjustments instead. The dependent variables are computed as the adjusted or unadjusted difference in the log of one plus the number of citations to all patents of a firm applied for in a given year and granted by 2006 in our data and the log of one plus the number of citations to the same set of successful patents of that firm in the same application year in the NBER 2006 dataset. Restricting the successful patents from our data

[7] We do not use for the analysis in this section firms without any patenting activity between 1976 and 2006.

to only those that are granted by 2006 allows for comparison with successful patents in the NBER 2006 data. Logs are taken to account for skewness in citation activity.

In specifications covering both the tables, we also iteratively employ time, technology class, industry, and firm fixed effects to account for characteristics that might be driving the biases in patent and citation activity. We include a host of firm-level variables (described in detail in Online Appendix D) to assess how the patent and citation bias might be related to these characteristics. This allows us to explore if—conditional on adjustments for time and technology class and accounting for time, technology, industry, and firm fixed effects—these biases are orthogonal to firm characteristics. Put another way, can the measurement error in patent and citation counts at the firm level due to truncation issues be called "classical" measurement error? If the error is classical, we may not have to worry about such biases confounding inferences about our explanatory variables in many cases.

Several facts emerge from the analyses in Tables 1 and 2, as well as the additional regressions reported in Online Appendix E. First, we see that various firm-level measures—firm size, R&D to sales, market-to-book ratio, cash to total assets, leverage, and spread—are statistically related to *both* patent and citation bias in most specifications. For instance, larger firms, as well as those with high market-to-book ratios, invariably have greater patent and citation bias, an effect that is statistically significant at the 5% confidence level in all reported regressions. Similarly, the R&D-to-sales, leverage, and cash to total assets ratios are also, by-and-large, strongly positively related to both patent and citation bias. Spread is also by-and-large strongly negatively related to these biases. ROA also tends to explain patent bias, with higher return firms exhibiting more bias. Importantly, these relationships exist even after we control for time, technology class, and industry fixed effects. In addition, these patterns survive the most stringent

specification with time and firm fixed effects. These results suggest that our findings exist even when we exploit within firm variation and therefore are not simply driven by the changing composition of firms over time.

One can rationalize the firm-level relationships in these specifications. Larger firms, as well as those that spend heavily on R&D, might produce more complex patents that require a longer time to be approved. Consequently, these patents might take longer to be granted and, for those that are granted, to accrue citations, leading to a positive patent and citation bias. One can make analogous arguments for why higher cash-to-assets (greater financial strength) and lower spreads (higher liquidity) might be related to these biases.

Second, the economic magnitudes of these relationships suggest that a large portion of both unadjusted patent and citation biases is explained by firm-level variables. For instance, a one standard deviation change in log size is associated with a 0.05 standard deviation change in unadjusted patent bias (column 2, Table 1). A similar change in log size is also associated with a 0.12 standard deviation change in unadjusted citation bias (column 2, Table 2). Similarly, a one standard deviation change in log R&D-to-sales ratio is related to about 0.03 standard deviation change in unadjusted patent bias (column 2, Table 1) and 0.5 standard deviation change in unadjusted citation bias (column 2, Table 2).

Third, the economic magnitudes we discussed above are similar when we use adjusted patent and citation biases as dependent variables instead. For instance, a one standard deviation change in log size is related to about 0.01 standard deviation change in adjusted patent bias (column 5, Table 1) and about 0.08 standard deviation change in adjusted citation bias (column 5, Table 2). Moreover, the effects are similar when we consider the most stringent specifications with firm fixed effects. In particular, a *within firm* one standard deviation change in size is related to

about a *within firm* 0.02 standard deviation change in adjusted patent bias (column 6, Table 1) and 0.03 standard deviation change in adjusted citation bias (column 6, Table 2). It is worth noting that while the magnitudes are larger in the case of citation bias, they are reasonably large for patent bias as well. This is the case for both unadjusted and adjusted variables being employed in the analysis. Recall, that we had picked the most stringent patent and citation level adjustment for presentation.

Fourth, the log of the total number of granted patents in the same technology class as the modal technology class of the firm's patents (log (Patents in Technology Class)) and the log of total number of citations to granted patents in the same technology class as the modal technology class of the firm's patents (log (Citations in Technology Class)) are strongly positively related to both patent and citation bias. Tables 1 and 2 reveal that this strong relationship exists across specifications, including those with industry, technology class, time, and firm fixed effects. In addition, there is weaker evidence that citation and patent biases are negatively related to log of the total number of granted patents in the same state as the modal state of the assignee on firm's patents (log (Patents in State)) and log of the total number of citations to granted patents in the same state as the modal state of the assignee on firm's patents (log (Citations in State)). This is certainly the case when we employ specifications with firm and time fixed effects (columns 3 and 6 of both Tables 1 and 2). These findings underscore the importance of technology class and region on inferences at the firm level. The different signs of the relationships further confirm the complex nature of these biases as they relate to technologies and regions, as well as the inability of firm fixed effects, time fixed effects, and popular patent-level adjustments to alleviate them.

Finally, we conduct in unreported analyses several additional tests that confirm the reliability of inferences derived above. For instance, citation and patent biases are more strongly

related to firm characteristics in patents granted towards the end of the sample. This is assuring because, as discussed earlier, younger patents are ones where one expects biases to be systematically related to firm characteristics.

We reach similar conclusions when using the new method for adjustment of truncation proposed by Jaffe and de Rassenfosse (2017). The authors suggest comparing patenting or citation activity with the "group of patents" to which the patent of interest belongs. Therefore, the group of patents considered for this adjustment, following this new approach, is all those granted to all the publicly traded firms. This differs from the earlier adjustments, where the comparison set is the entire population of patents. In interest of brevity, we present this analysis in Online Appendix E. As Table E2 shows, when aggregated at the firm level, these biases are still related to firm characteristics, after applying these adjustments. As before, variables such as market-to-book ratio, R&D expenditures, and leverage predict variation in patent and citation biases at the firm level.

Taken together, this analysis shows that the patent-level truncation problems create complex biases both in terms of patents granted and citations, when patents are aggregated at the firm level. These biases are not addressed by the usual adjustment methods in the literature. Adding fixed effects at the firm, industry, and year level also are not sufficient. Substantial patent and citation bias remains, which is systematically related to firm characteristics such as size, R&D intensity, leverage, cash-to-assets ratio, and spread. It is also related to technological and regional characteristics of firm's patenting. As a result, several inferences that researchers might attribute to a phenomenon that they are studying may instead be driven by the bias—i.e., by non-classical measurement errors.

4. A Robustness Checklist and a Way Forward

Our paper has highlighted how patent and (particularly) citation biases can be correlated with key firm-level characteristics that are of interest to researchers. While the basic issues confronting users of patent data are similar to ones that those analyzing almost any contemporaneous database face, the dramatic changes in the direction and location of technological innovation (and patenting practice) over recent decades have led to biases in patent-based analyses.

As we have highlighted, there is no single adjustment method that can straightforwardly address these issues. The dynamic nature of technological change, shifts in patent policy, and endogenous firm responses means that any formulaic set of adjustments would soon be out-of-date.

In this section, we suggest two sets of remedies to these problems: a robustness checklist and a more involved longer-term remedy using machine learning.

4.A A Checklist

As we have highlighted above, patent data poses significant and often poorly understood challenges to scholars in finance and related fields. In particular, the interaction of the truncation of patent data and the changing composition of inventors in a region or sector can lead to significant distortions that may affect firms with particular characteristics, such as those of a large size or with low book-to-market ratios. These distortions can lead to problematic inferences, particularly when studying reasonably proximate events (or more precisely, those reasonably proximate to the end of the patent data-set).

To avoid these pitfalls, we would suggest that researchers subject their analysis to a checklist of robustness tests (Table 3). By applying these additional robustness tests, they can assure themselves and their readers that many of the problems we documented have been avoided.

These checks are not intended to create a series of impossible hurdles for researchers that seek to use patent data. Rather, they are intended as parallel diagnostic approaches that should serve as a springboard for follow-on questions, if the answers are troubling. For instance, if in response to the evaluation suggested by item 5 (*"robustness with respect to firm characteristics"*), it is found that the results look very different with and without the inclusion of high book-to-market firms, it would be appropriate to better understand the roots of this disparity. The researcher should explore to what extent the effect is seen throughout the sample, or if it is concentrated in the sample's final years. In the latter case, it might be appropriate to worry about the deleterious effects of biases, and explore the robustness of the results further. As a concrete example, consider the literature on the impact of state anti-takeover legislation on innovation that has suggested widely varying patterns (for instance, Atanassov (2013), Becker-Blease (2011), and Chakraborty, Rzakhanova, and Sheikhb (2014)). Were these earlier works to have applied the tests articulated in item 4 (*"robustness with respect to geography"*) and item 5 (*"robustness with respect to firm characteristics"*) to their analyses, we believe, the authors would almost certainly have been far more guarded in their conclusions regarding the relationship between anti-takeover legislation and innovation. Similar arguments apply to the literatures that connect banking deregulation and policies that altered liquidity in the stock market to innovation, among others.

The first set of robustness tests asks researchers to focus on biases due to differences/changes in composition of their samples. As we have discussed, spurious results regarding the impact of policy changes or firm decisions may result when changes in the

composition of sub-samples are ignored. Our recommendations in the first few tests in the checklist are therefore to:

- Provide results on how the patent and citation bias are related to the policy or firm choices being studied:
  - The first step here is to compute patent and citation biases using patents granted to the same firms and applied for during the same time period, employing different versions of patent data, much like our analysis in Section 3.
  - Our recommendation is to compute these biases after adjusting for technology and time effects at the patent level, since this seems to be the most stringent patent-level correction.
  - These can then be related to the key policy changes under study, to see whether the measurement errors are correlated with the variables of interest.
- Provide estimates of patent and citation changes in response to the policy or firm decision being studied for different subsamples across:
  - Different time periods: It is important in particular to assess the role of few years towards the end of the sample. It is particularly critical to do this test when evaluating policies or choices where much of the activity of interest lies within a decade of the end of the database.
  - Different technology classes: It is important in particular to assess the role of heavy patenting technology classes highlighted in Section 2. Again, the acceleration of activity in recent decades may play havoc with inference otherwise.

- Different states: It is important in particular to assess the role of regions that have experienced a recent acceleration in patenting, such as California and Massachusetts, as also highlighted in Section 2.
- Different classes of firms: It is important in particular to assess the role of firms that have experienced a particular acceleration of patents, such as those with a high market-to-book value.

The final set of tests explore the robustness of the results to the more general limitations of the patent data, as discussed in Online Appendix H. These include exploring the potential distortions introduced by:

- The exit of firms.
- The inability to match exactly patents to listed firms.
- The potential for strategic behavior in patent assignment and citation.

To help researchers, we have captured these considerations in the form of eight key checklist items, which we list in Table 3.

Again, we should emphasize that as with any checklist, it must be utilized by researchers, editors, and referees with judgement. Few papers will be perfect along all dimensions: that does not mean that these papers should not be published. Similarly, some papers may look good along seven criteria but deeply flawed on the eighth: these may merit rejection.

4.B The Application of Machine Learning

We next explore the application of machine learning to address concerns about patent and citation bias. In particular, our analysis in the paper suggests that the traditional adjustments discussed above by themselves are not able to systematically address the biases in patents and

citations at the firm level. Moreover, these biases tend to be correlated with firm-level characteristics, rendering analysis and inferences using firm-level data open to erroneous conclusions.

The conventional adjustments are targeted to think about various biases at the patent level. We now ask if using machine learning—exploiting a richer set of data at the firm level—might help in reducing these biases. In this section, we are not trying to develop the best, state-of-the-art machine learning model to address this problem. Rather, our approach ambition is more modest. We assess if using even basic ML methods might do "better" than using conventional patent-level adjustments when doing firm-level analyses. As such, that analysis is designed to be more indicative of the potential of this approach more than anything else.

We provide a short summary here; a more technical account is in Online Appendix I. The interested reader is also referred to the relevant code and underlying data, which are posted at https://github.com/YuanHBS/Patent-Counts-and-Citation-Prediction.

We undertake two sets of machine learning predictions. The first seeks to address the patent-level bias at the firm level and the second one targets citation bias:

- In the first analysis, we focus on patent bias. For each firm that has a Compustat identifier in the 2006 NBER database, we attempt to predict the number of patents that will be ultimately granted to each firm in each filing year and HJT technology class between 2002 and 2006. We then compare the predicted values to the actual granted patents in these filing years and classes recorded as of the end of 2012. To do this, we proceed as follows:
    - We use patents applied for between 1976 and 2001 as the training data for the ML models. For this set, we determine the best predictors of the number of patents applied for in each year and awarded through 2006. To make the predictions, we

use several explanatory variables: the counts of patent applications made by the firm in the prior six years and awarded by the end of 2006, broken down the patent category (which of the six HJT classes that the patent has a primary assignment to), and a variety of financial firm-level variables. We use six popular ML models to make the predictions. (Online Appendix I presents other variants of these models.) [8]

  - Using the relationships established using the training data, we predict the number of patents that will be granted to each firm in each year and class between 2002 and 2006.
  - We then compare the predicted values to the actual granted patents, recorded as of the end of 2012.

- In the second analysis, we focus on citation bias. We seek to predict the total citations received by each patent in the first ten years after ultimately granted patents at the firm level every year. We use patents applied for between 1976 and 1992 as the training data for the ML models.[9] We then compare the predicted values to the actual citations—for patents filed between 1993 and 2002—recorded as of the end of 2012.
  - For this set, we determine the best predictors of the citations received within ten years of issue by patents applied for in a given year. To make the predictions, we

[8] One subtle issue is that the model must account not only for the relationship between patent activity in the prior six years and that in the year under examination, but also the changing impact of patent bias. As we approach 2001, not only are the awards in the year under examination subject to patent bias, but so are those in the six years prior as well.

[9] The assumption underlying this choice was that the typical patent applied for in 1992 would have issued by 1996, and thus we would be able to accurately count the number of citations received in its first ten years by the end of 2006. Those received in subsequent years would not necessarily have a full ten years to be cited.

use several explanatory variables: the counts of patent applications made by the firm in the past six years and awarded by the end of 2006, broken down the patent category (which of the six HJT classes that the patent has a primary assignment to), the same financial firm-level variables, and the predicted patent count in the earlier analysis.

- We then predict, based on these coefficients, the number of citations in the ten years after issue for patents applied for by a firm in the years from 1993 and 2002.
- We compare the predicted values to the actual citations recorded as of the end of 2012.

Having derived the predicted patent and citation counts from the ML model, we then need to assess their performance. To do this, we create alternative estimates of patenting and citation activity that do not involve machine learning. We will compare accuracy of the various predictions, in order to assess whether the machine learning estimate is better or worse than these "less sophisticated" estimates.

The first, which we term the raw benchmark, is the actual patents or citations in the decade after the patent issues activity in the 2006 NBER dataset. Thus, we use the number of patents applied for by each firm in each year and patent class between 2002 and 2006 and awarded by the end of 2006 on the one hand, and the number of citations that patents applied for by a firm in each year from 1993 to 2002 and patent class had received as of the end of 2006.

But we know that these predicted values are going to be undercounts, as they do not include any activity (additional patents granted or citations received for patents filed before 2006) between 2007 and 2012. For instance, we know that while a firm will have a rather modest share of the patents that it applied for 2002 remaining unissued by December 2006, many more of the patents

it applied for in 2005 will not have issued by the end of 2006. We thus developed a second prediction, which we term the time-adjusted benchmark. This alternative measure makes extrapolations, akin to the time-adjustment methodology described above, for both the patent and citation counts. The precise formula for these adjustments is described in the appendix, but essentially it involves adjusting (scaling) the patents and citation counts upward based on overall historical patterns.

But the time-adjusted benchmark is not the best adjustment we could do, as it does not reflect the differences in the evolution of patenting across technology classes discussed above. We thus developed a third benchmark, which we term the enhanced time-and-technology adjusted benchmark. The precise formula for these adjustments is described in the appendix, but essentially it involves further tuning the patents and citation counts by incorporating technology class patterns. For citations, we also employed the quasi-structural methodology, based on the methodology described in Appendix B.

We then compared the output of the ML model with the raw benchmark, the time-adjusted benchmark, and the enhanced time-and-technology adjusted benchmark for patents, and with these three benchmarks and the quasi-structural one for citations. We used two primary statistical approaches to make these comparisons. In the root mean square error (RMSE) approach, the smaller the test statistic, the better the prediction. For the R squared and Pearson product-moment approach, the closer the test statistic is to one, the better. The RMSE and R squared approaches are the most popular ways to evaluate machine learning predictions, while the correlation is considered as an auxiliary metric.

As can be observed from Tables 4 and 5, the ML approaches perform strongly. All our six ML models that use firm-level information consistently predict the patent counts better than the

raw, time-adjusted, and enhanced time- and technology-adjusted data. Similarly, our ML models using firm-level information consistently predict the patent citations better than the benchmarks. In Appendix I, we report the results on several additional analyses using modified approaches to estimations in several additional tables. These robustness checks also perform considerably better than the traditional methods of addressing patent and citation bias. These results strongly suggest that using machine learning methods that rely on firm information have considerable power in this setting.

Overall, we have developed and incorporated an analysis that uses machine learning to generate adjustments that can perform considerably better than traditional adjustments used in the literature. We illustrated that adjustments generated by a variety of machine learning-based models using firm-level information–in addition to patent-specific information that is employed for popular "patent-based" adjustments in the literature–minimizes patent and citation bias at the firm level. While the goal of this analysis is simply to be indicative of the potential of this methodology, it underscores the importance of using firm-level information when adjusting for patent and citation bias and points to a direction to resolving these challenges. Incorporating more advanced ML methodologies might further increase the prediction accuracy and is a promising future research direction.

4.C Final Thoughts

We conclude with a topic for future research. One enduring point is the limitations of the existing systems for addressing truncation issues, especially in light of the systematic biases that we highlight above. We are struck by the power of even a parsimonious set of machine learning methodologies. In particular, it suggests that incorporating firm-level information appears to

contribute considerable information to the inference process. Assessing what other firm-level information could be used for efficient prediction of patents and citations at firm level (and therefore for developing firm-level adjustments) remains a fruitful area of research.

References

Atanassov, J. 2013. Do Hostile Takeovers Stifle Innovation? Evidence from Antitakeover Legislation and Corporate Patenting. *Journal of Finance* 68:1097–1131.

Becker-Blease, J. 2011. Governance and Innovation. *Journal of Corporate Finance* 17:947–958.

Bena, J., M. Ferreira, P. Matos, and P. Pires. 2017. Are Foreign Investors Locusts? The Long-Term Effects of Foreign Institutional Ownership. *Journal of Financial Economics* 126:122-146.

Bessen, J. 2009. NBER PDP Project User Documentation: Matching Patent Data to Compustat Firms. Unpublished working paper, Boston University.

Chakraborty, A., Z. Rzakhanova, and S. Sheikhb. 2014. Antitakeover Provisions, Managerial Entrenchment and Firm Innovation. *Journal of Economics and Business* 72:30–43.

Chemmanuer, T., and X. Tian. 2018. Anti-Takeover Provisions, Innovation, and Firm Value: A Regression Discontinuity Analysis? *Journal of Financial and Quantitative Analysis* 53:1163-1194.

Dass, N., V. Nanda, and S. Xiao. 2017. Truncation Bias Corrections in Patent Data: Implications for Recent Research on Innovation. *Journal of Corporate Finance* 44:353-374

Dyck, A., A. Morse, and L. Zingales. 2010. Who Blows the Whistle on Corporate Fraud? *Journal of Finance* 65:2213-53.

Hall, B., A. Jaffe, and M. Trajtenberg. 2001. The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Working Paper 8498, National Bureau of Economic Research.

Jaffe, A., and G. de Rassenfosse. 2017. Patent Citation Data in Social Science Research: Overview and Best Practices. *Journal of the Association for Information Science and Technology* 68:1360-74.

Jaffe, A., and M. Trajtenberg. 2002. *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, Cambridge, MA: MIT Press.

Karpoff, J., A. Koester, D. Lee, and G. Martin. 2014. Database Challenges in Financial Misconduct Research. Research Paper No. 2012-15, Georgetown McDonough School of Business.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman. 2017. Technological Innovation, Resource Allocation, and Growth. *Quarterly Journal of Economics* 132:665–712.

Kolev, J. 2013. Credit Constraints and their Impact on Innovation: Evidence from Venture Capital Exits. Unpublished Working Paper, Harvard University.

Lerner, J., M. Sorensen, and P. Strömberg. 2011. Private Equity and Long-Run Investment: The Case of Innovation. *Journal of Finance* 66:445-477.

Li, G., R. Lai, A. D'Amour, D. Doolin, Y. Sun, V. Torvik, A. Yu, and L. Fleming. 2014. Disambiguation and Co-Authorship Networks of the U.S. Patent Inventor Database (1975–2010). *Research Policy* 43:941-955.

Merges, R. 1992. *Patent Law and Policy: Cases and Materials*. Charlottesville, Virginia: Michie Company.

Moser, P., and A. Voena. 2012. Compulsory Licensing: Evidence from the Trading with the Enemy Act. *American Economic Review* 102:396-427.

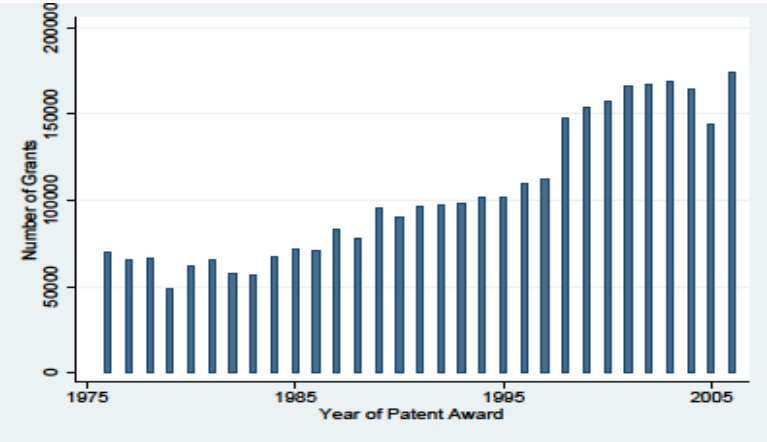
Nicholas, T. 2008. Does Innovation Cause Stock Market Runups? Evidence from the Great Crash. *American Economic Review* 98:1370-96.

Seru, A. 2014. Firm Boundaries Matter: Evidence from Conglomerates and R&D Activity. *Journal of Financial Economics* 111:381-405.

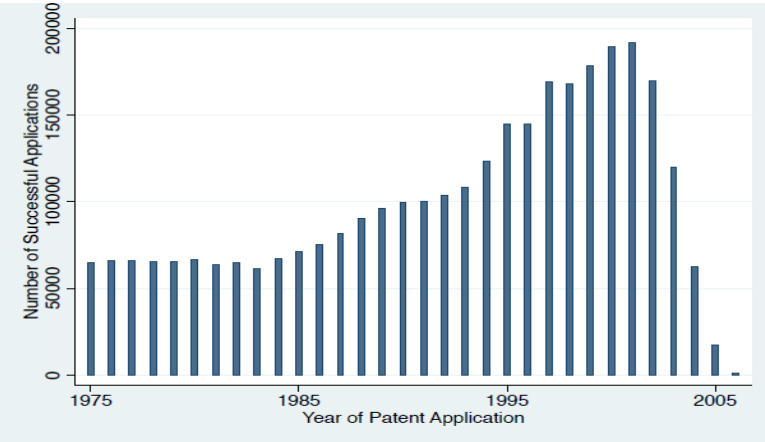
Thoma, G., S. Torrisi, A. Gambardella, D. Guellec, B. Hall, and D. Harhoff. 2010. Methods and Software for the Harmonization and Integration of Datasets: A Test Based on IP-Related Data and Accounting Databases with a Large Panel of Companies at the Worldwide Level. Unpublished Working Paper.

Figure 1: U.S. Grants, Successful Patent Applications, Total Patent Applications, and Citations per Patent

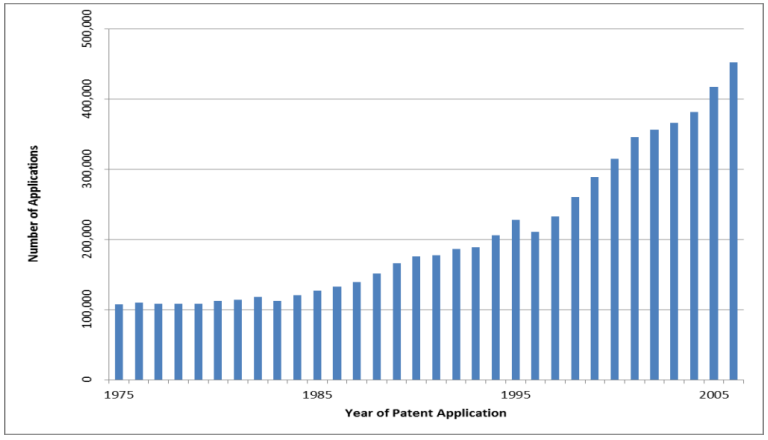
The figure shows the number of grants (Panel A), the number of successful applications (Panel B), and the citations per patent (Panel D) from 1975 through 2006 in the NBER 2006 patent database. The total number of applications in Panel C is defined by the number of successful and unsuccessful patent applications and is from the USPTO. Panel A highlights the three-fold increase in grants, while Panel C shows how total applications increased four-fold. Panels B and D highlight the truncation of the NBER database, which only reports the patents issued by the end of 2006, and saw citations peak at 1985 before falling rapidly. Source: NBER 2006 patent dataset and the USPTO website.



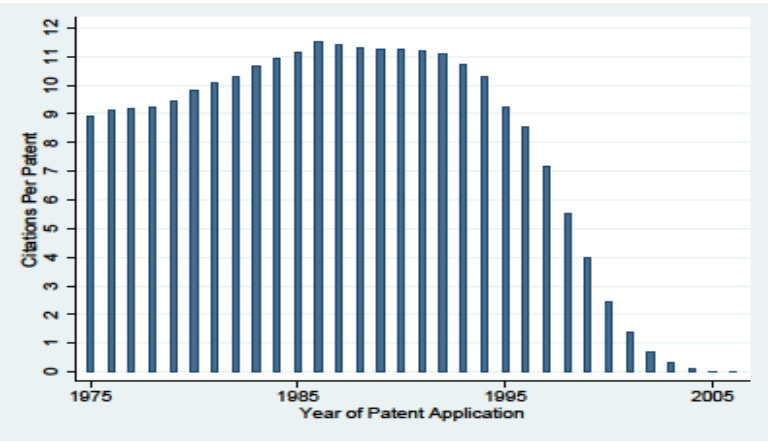
(a) U.S. patent awards over time, by award year



(b) U.S. successful patent applications over time, by application year



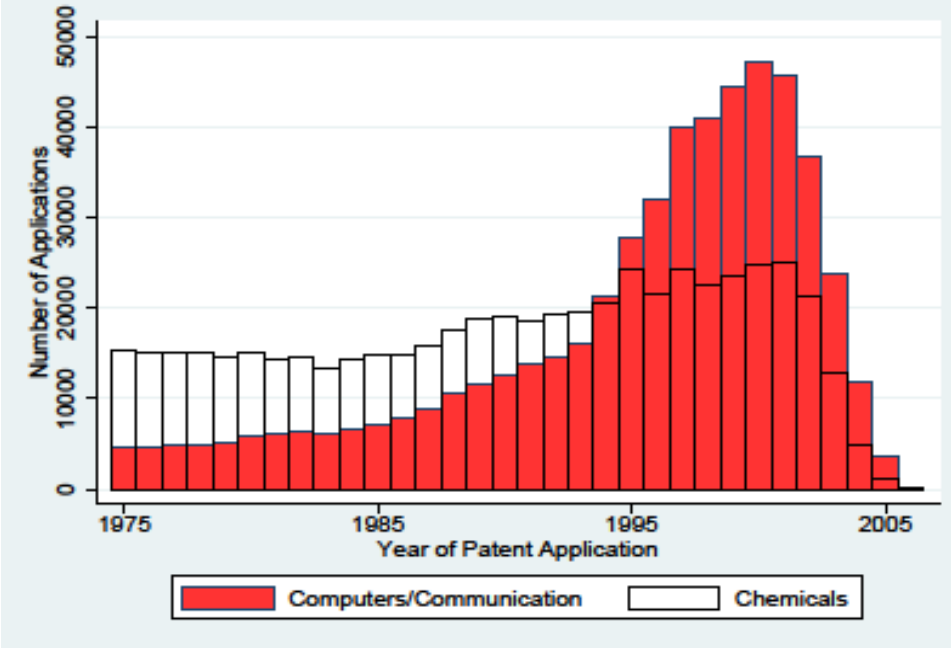
(c) Actual patent applications (successful and unsuccessful), by application year



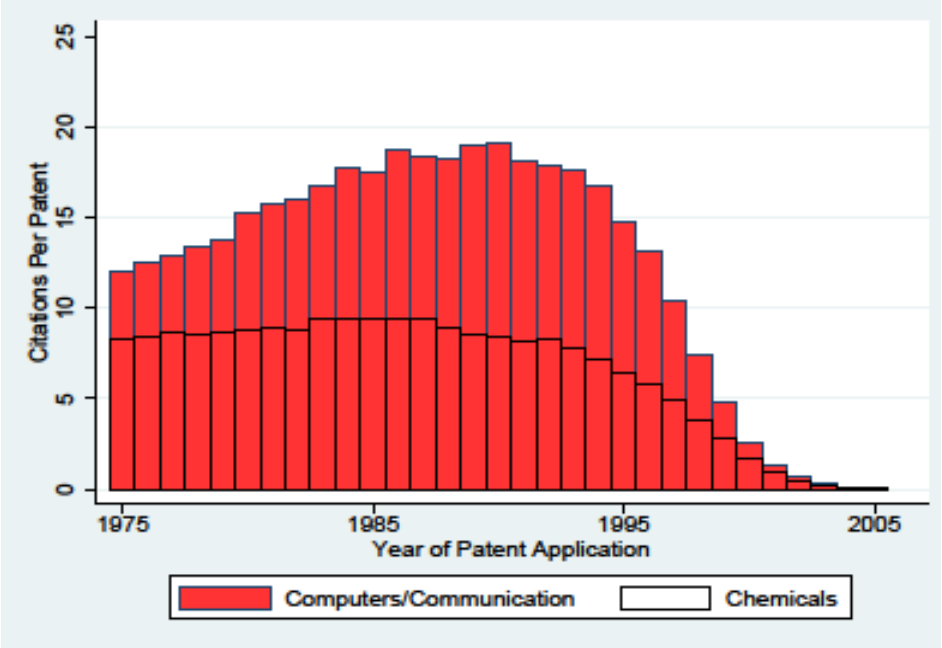
(d) Citations per patent over time, by application year

Figure 2: U.S. Successful Applications and Citation Comparisons, by Industry

Panel A graphs the number of successful patent applications in the "computers and communications" and "chemicals" HJT subcategories by the year of the application. Panel B plots the citations per patent for patents in the "computers/communication" and "chemicals" HJT subcategories by the year of patent application. The two graphs make clear that computers experienced a dramatic run-up in patenting activity in the 1980s to early 2000s and that there are dramatic differences in citation rates across technologies. These figures highlight the heterogeneity in truncation bias in successful applications and citations across various industry subcategories. Source: NBER 2006 patent dataset.



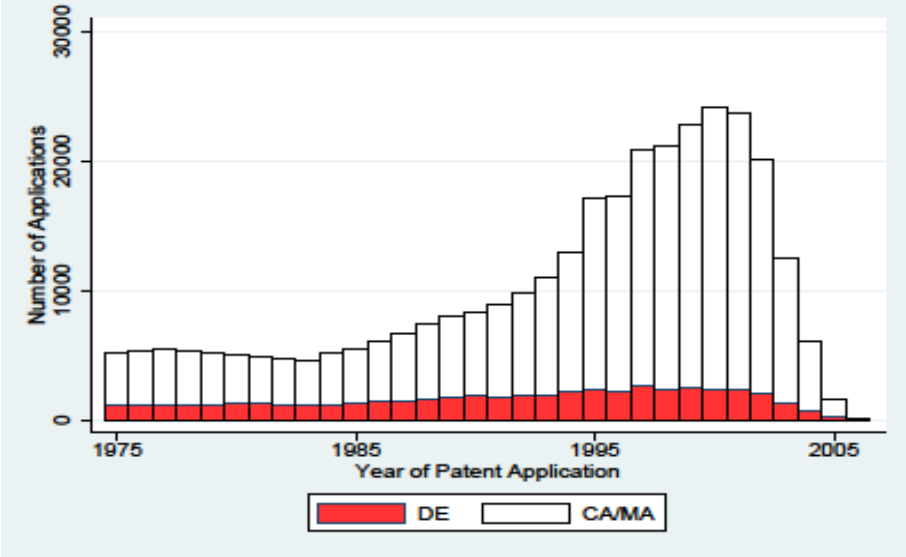
(a) U.S. successful patent applications, chemicals versus computers



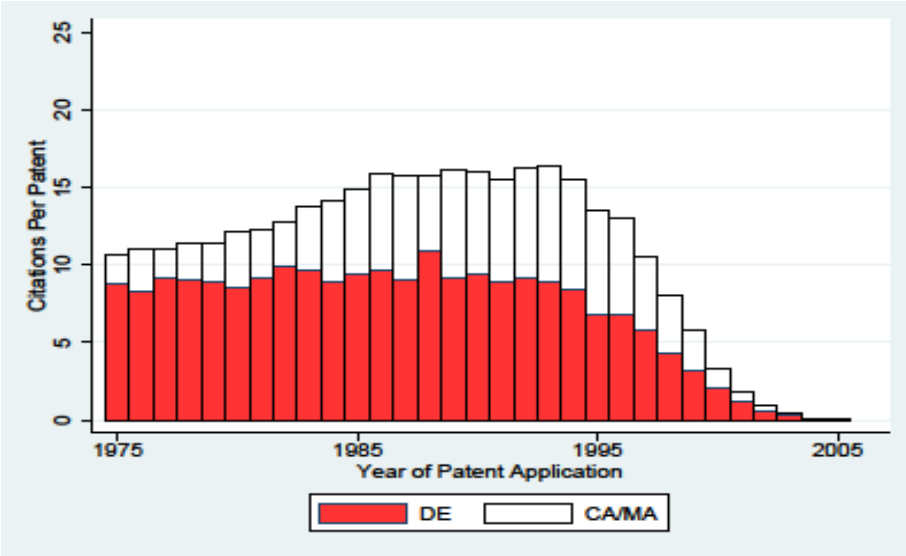
(b) Citations by patent application year, chemicals vs. computers

Figure 3: U.S. Successful Applications and Citation Comparisons, by State of Assignee

Panel A presents the number of successful patent applications over time by the state of assignee, with Delaware (DE) in red and California and Massachusetts (CA/MA) in white; Panel B the number of citations per patent by state of assignee. The graph shows that growth in patenting is far from uniform geographically, as there was a dramatic increase in successful patents from assignees in CA and MA over time, whereas patents stayed more or less the same in DE. Citations per patent are also geographically different: patents by assignees from CA and MA were far more likely to get cited than those from DE. These comparisons suggest that any naïve correction for the truncation of patents and citations that does not account for such dramatic regional differences may lead to incorrect inferences. Source: NBER 2006 patent dataset.



(a) U.S. successful patent applications, DE versus CA and MA



(b) Citations by patent application year, DE versus CA and MA

Figure 4A: Distribution of Firm Patent Bias (Unadjusted and Adjusted) over Time

This figure presents the distribution of patent bias aggregated at the year level (upper graph) and on a mean firm-year level (lower graph) for patents granted to public firms in each year between 1976 and 2006. To compute the unadjusted patent bias for each firm-year, we compare the number of patents for each firm filed in each application year in our data (thus, which have been granted by 2012) and in the NBER 2006 dataset (i.e., granted by 2006). The adjustments use the time fixed effect and the time and technology class fixed effect methodologies, with details discussed in the text and Online Appendix B. Sources: NBER 2006 patent and our datasets.

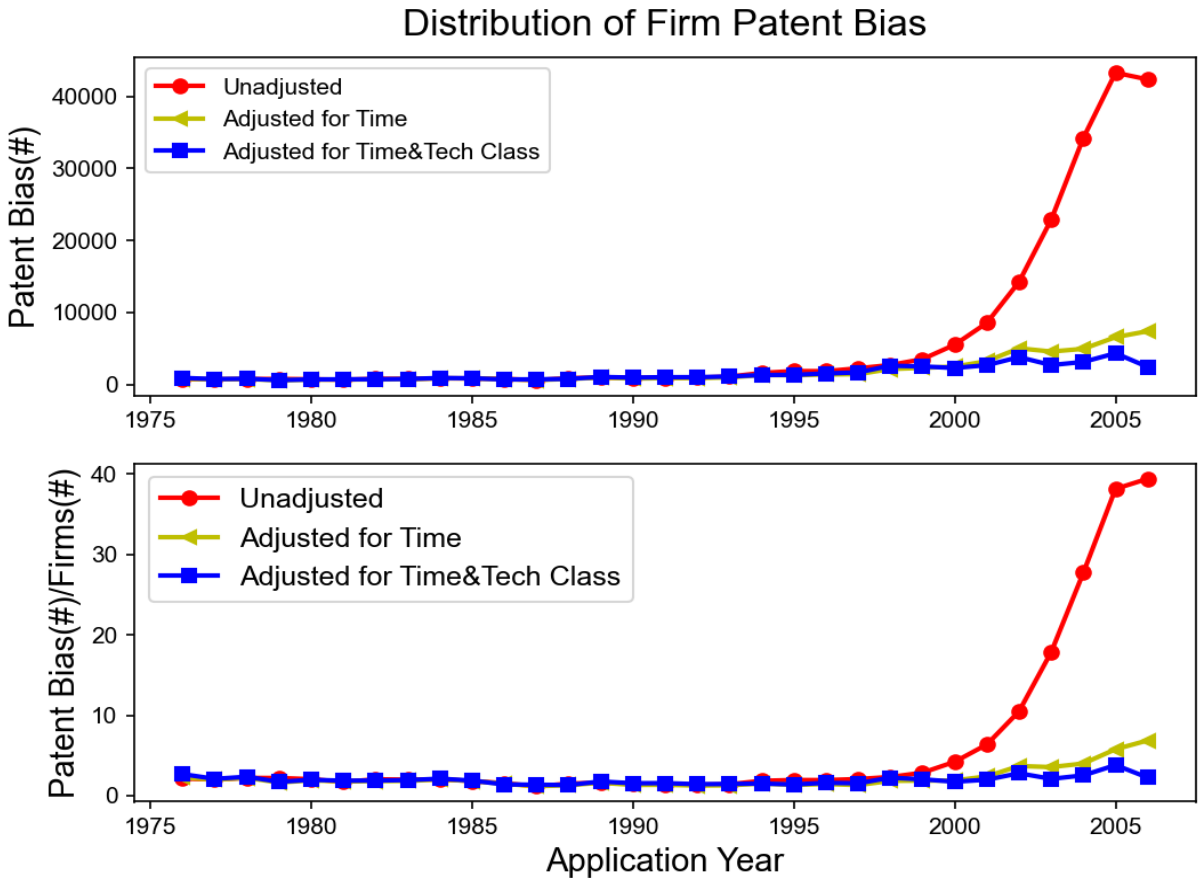


Figure 4B: Distribution of Firm Citation Bias (Unadjusted and Adjusted) over Time

This figure presents the distribution of citation bias aggregated at the year level (upper graph) and on a mean firm-year level (lower graph) for patents granted to public firms from 1976 through 2006. To compute the unadjusted citation bias for each firm-year, we compare the number of citations to all the patents for each firm filed in each application year in our data (i.e., citations in patents granted by 2012 to applications filed by a firm in a given year and granted by 2006) and in the NBER 2006 dataset (i.e., citations in patents granted by 2006 to applications filed by a firm in a given year and granted by 2006). The adjustments use the time fixed effect methodology, the time and technology class fixed effect methodology, and the citing year effect adjustment using the quasi-structural method, with details discussed in the text and Online Appendix B. The lines for the time fixed effect methodology and the time and technology class fixed effect methodology are almost superimposed due to the scale. Sources: NBER 2006 patent and our datasets.

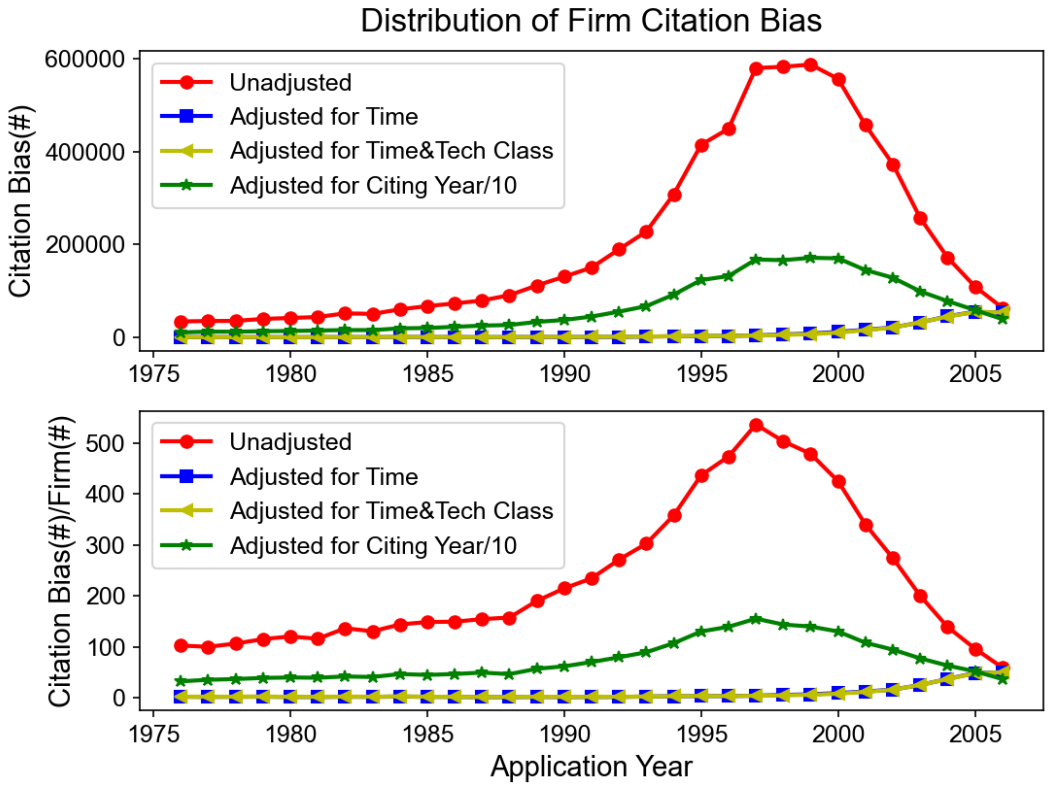
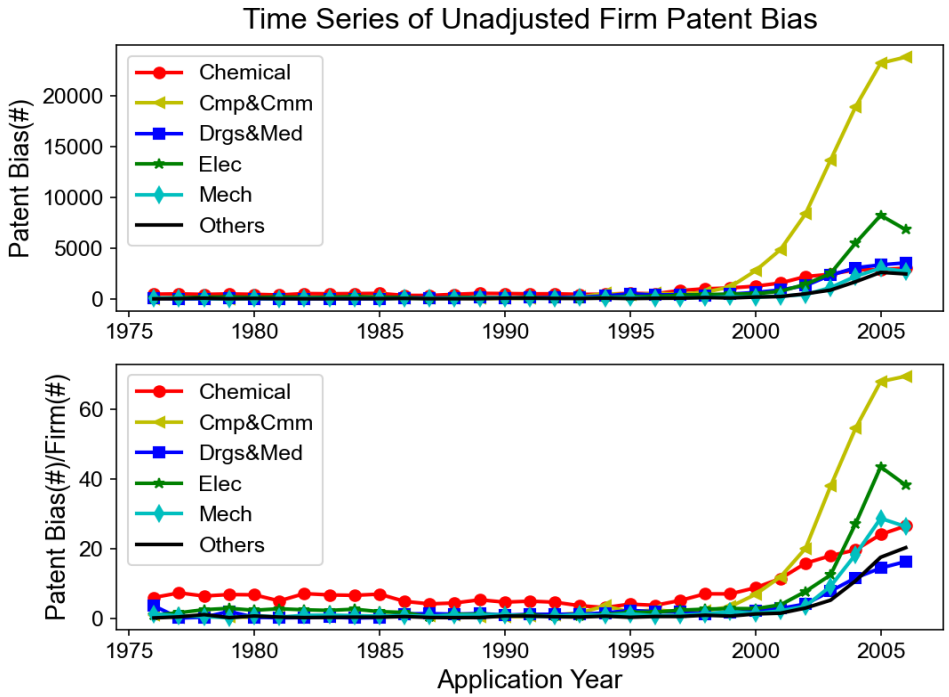
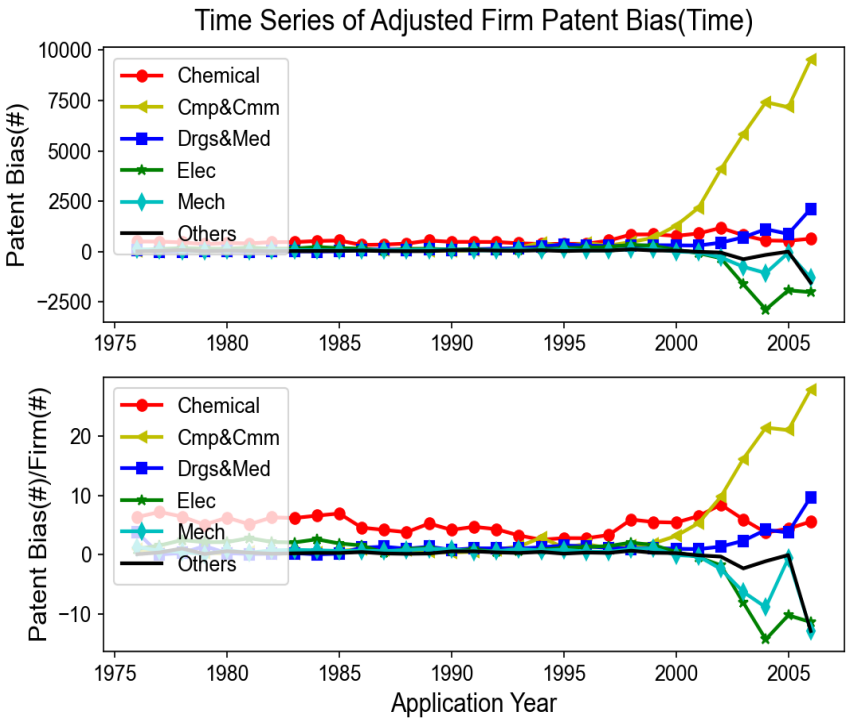


Figure 5: Firm Patent Bias (Unadjusted and Adjusted) across HJT Technology Classes

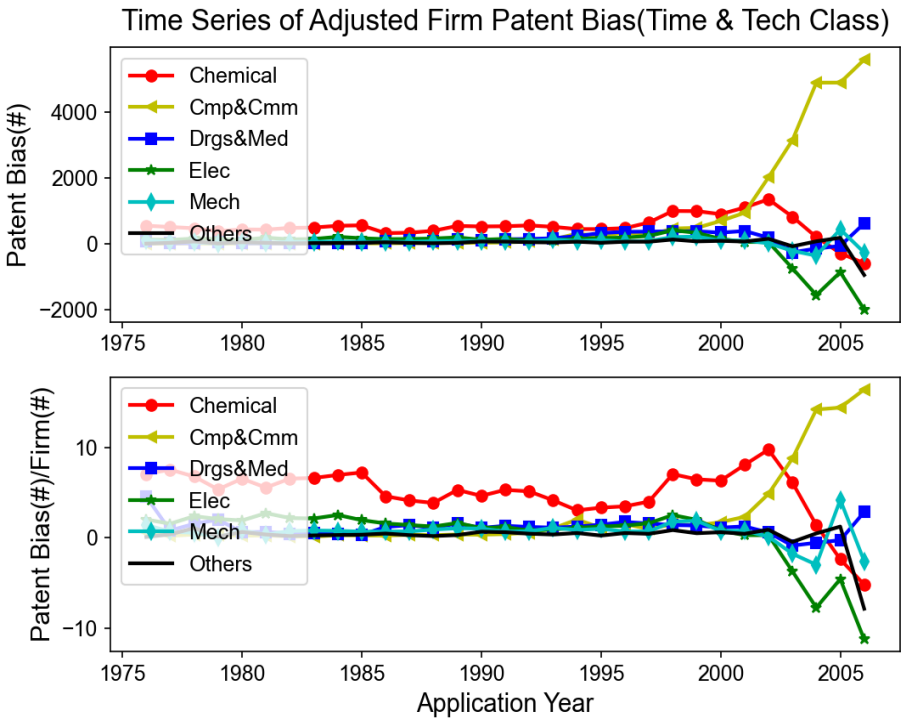
This figure presents the distribution of patent bias aggregated at the year level (upper graph in each panel) and on a mean firm-year level (lower graph) for patents granted to public firms in each year between 1976 and 2006 in different HJT technology classes. To compute the unadjusted patent bias for each firm-year, we compare the number of patents for each firm filed in each application year in our data (thus, which have been granted by 2012) and in the NBER 2006 dataset (i.e., granted by 2006). A firm is assigned to a particular technology class in a given year based on the modal primary patent class of patents produced by that firm in that year, based on the U.S. patent classification system. The adjustments use the time fixed effect and the time and technology class fixed effect methodologies, with details discussed in the text and Online Appendix B. Sources: NBER 2006 patent and our datasets.



(a)


Time Series of Adjusted Firm Patent Bias(Time)
Patent Bias(#)
Chemical
Cmp&Cmm
Drgs&Med
Elec
Mech
Others
Patent Bias(#)/Firm(#)
Application Year

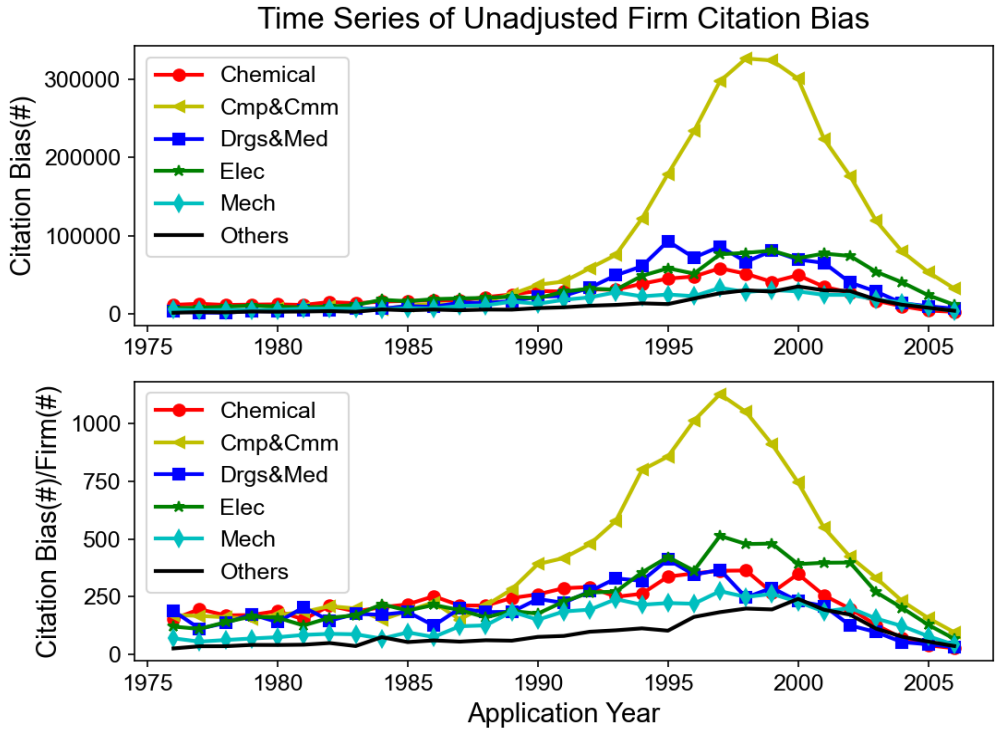
(b)


Time Series of Adjusted Firm Patent Bias(Time & Tech Class)
Patent Bias(#)
Chemical
Cmp&Cmm
Drgs&Med
Elec
Mech
Others
Patent Bias(#)/Firm(#)
Application Year

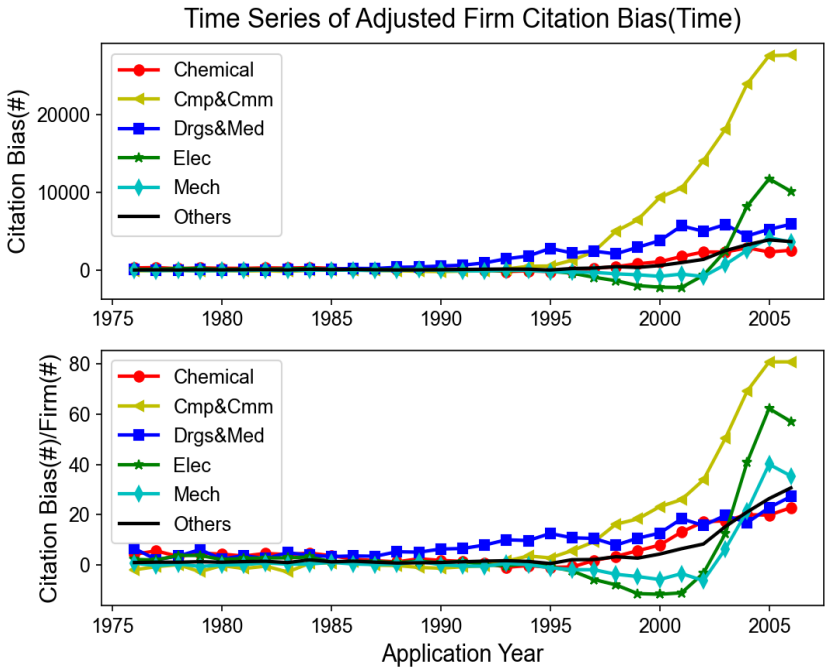
(c)

Figure 6: Firm Citation Bias (Unadjusted and Adjusted) across HJT Technology Classes

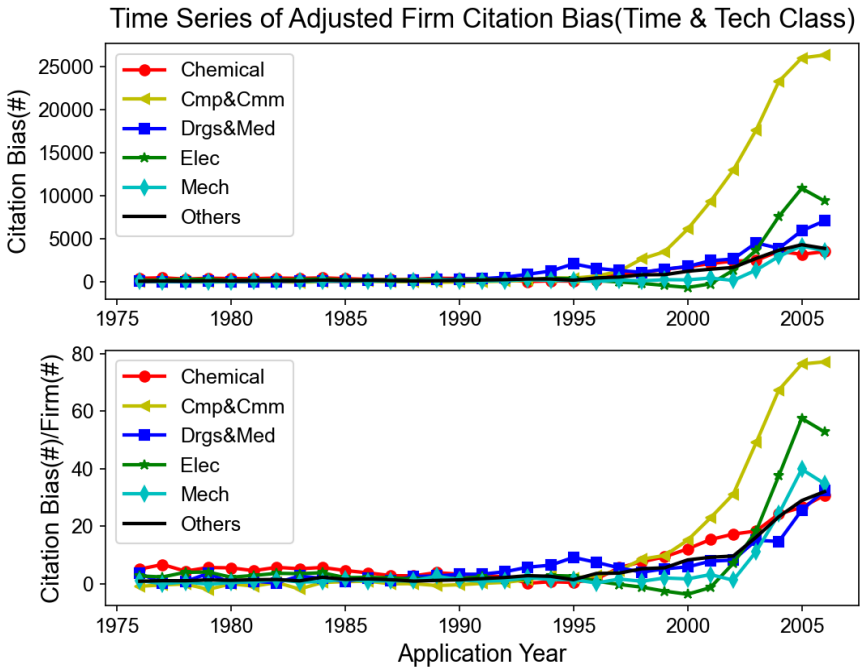
This figure presents the distribution of citation bias aggregated at the year level (upper graph in each panel) and on a mean firm-year level (lower graph) for patents granted to public firms from 1976 through 2006 in different HJT technology classes. To compute the unadjusted citation bias for each firm-year, we compare the number of citations to all the patents for each firm filed in each application year in our data (i.e., citations in patents granted by 2012 to applications filed by a firm in a given year and granted by 2006) and in the NBER 2006 dataset (i.e., citations in patents granted by 2006 to applications filed by a firm in a given year and granted by 2006). A firm is assigned to a particular technology class in a given year based on the modal primary patent class of patents produced by that firm in that year, based on the U.S. patent classification system. The adjustments use the time fixed effect methodology, the time and technology class fixed effect methodology, and the citing year effect adjustment using the quasi-structural method, with details discussed in the text and Online Appendix B. Sources: NBER 2006 patent and our datasets.



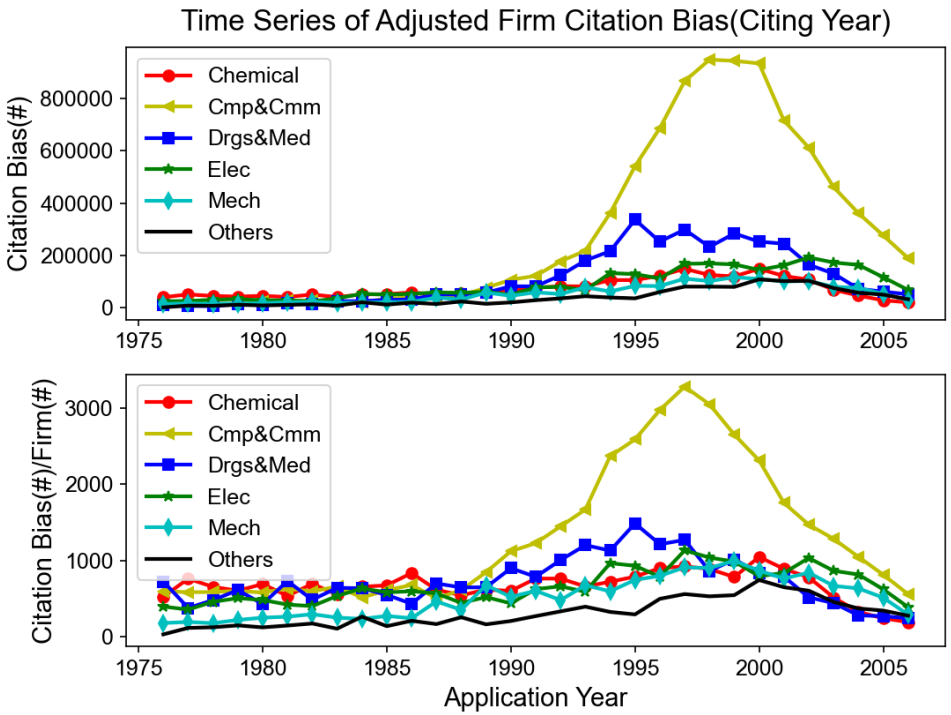
(a)


Time Series of Adjusted Firm Citation Bias(Time)
Citation Bias(#)
Citation Bias(#)/Firm(#)
Application Year
Chemical
Cmp&Cmm
Drgs&Med
Elec
Mech
Others

(b)


Time Series of Adjusted Firm Citation Bias(Time & Tech Class)
Citation Bias(#)
Citation Bias(#)/Firm(#)
Application Year
Chemical
Cmp&Cmm
Drgs&Med
Elec
Mech
Others

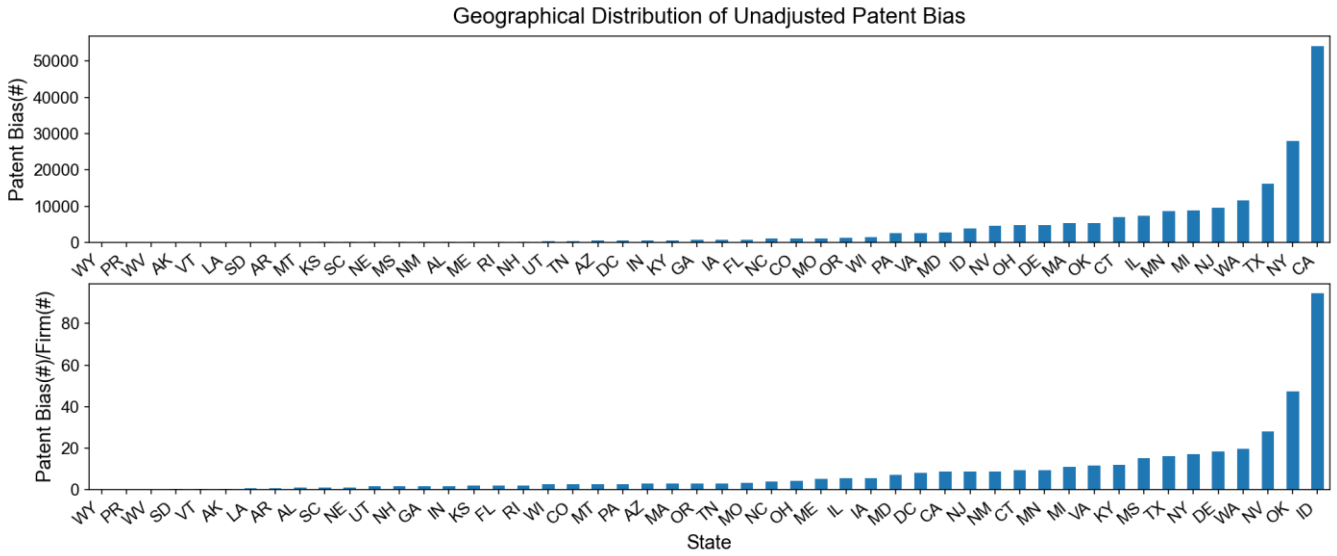
(c)


Time Series of Adjusted Firm Citation Bias(Citing Year)
Citation Bias(#)
800000
600000
400000
200000
0
Chemical
Cmp&Cmm
Drgs&Med
Elec
Mech
Others
1975
1980
1985
1990
1995
2000
2005
Citation Bias(#)/Firm(#)
3000
2000
1000
0
Chemical
Cmp&Cmm
Drgs&Med
Elec
Mech
Others
1975
1980
1985
1990
1995
2000
2005
Application Year

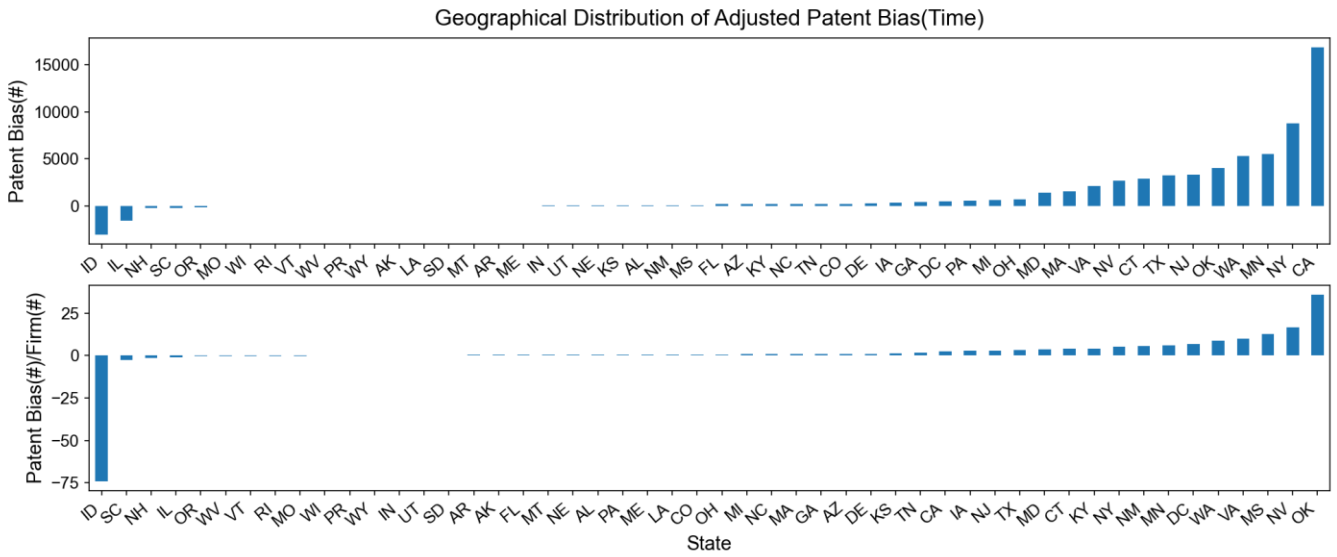
(d)

Figure 7: Firm Patent Bias (Unadjusted and Adjusted) across States

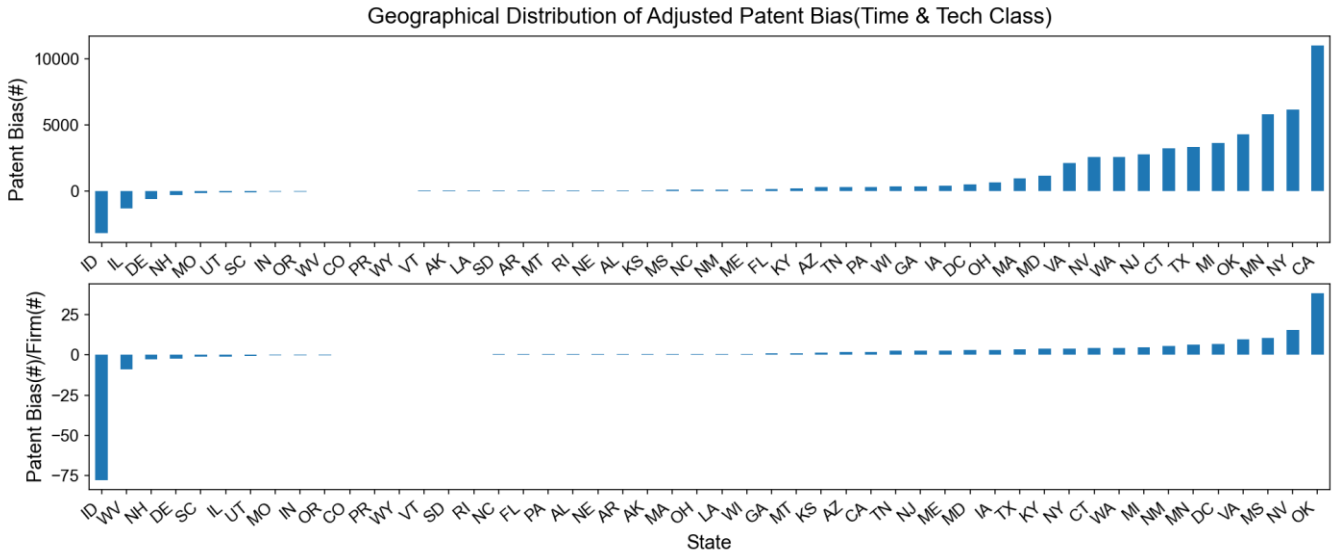
This figure presents the distribution of patent bias aggregated at the year level (upper graph in each panel) and on a mean firm-year level (lower graph) for patents granted to public firms in each year between 1976 and 2006 in different states. To compute the unadjusted patent bias for each firm-year, we compare the number of patents for each firm filed in each application year in our data (thus, which have been granted by 2012) and in the NBER 2006 dataset (i.e., granted by 2006). A firm is assigned to a particular state in a given year based on modal state of the assignees across patents granted to the firm at the time of the patent filing. The adjustments use the time fixed effect and the time and technology class fixed effect methodologies, with details discussed in the text and Online Appendix B. Sources: NBER 2006 patent and our datasets.



(a)


Geographical Distribution of Adjusted Patent Bias(Time)
Patent Bias(#)
15000
10000
5000
0
ID IL NH SC OR MO WI RI VT WV PR WY AK LA SD MT AR ME IN UT NE KS AL NM MS FL AZ KY NC TN CO DE IA GA DC PA MI OH MD MA VA NV CT TX NJ OK WA MN NY CA
Patent Bias(#)/Firm(#)
25
0
−25
−50
−75
ID SC NH IL OR WV VT RI MO WI PR WY IN UT SD AR AK FL MT NE AL PA ME LA CO OH MI NC MA GA AZ DE KS TN CA IA NJ TX MD CT KY NY NM MN DC WA VA MS NV OK
State

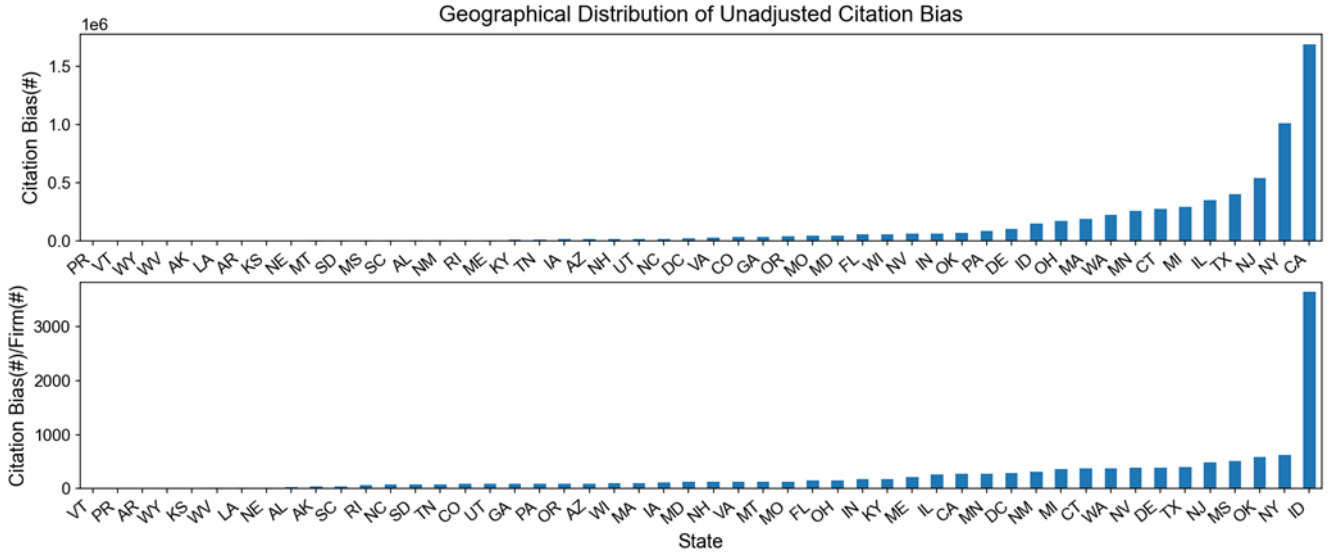
(b)


Geographical Distribution of Adjusted Patent Bias(Time & Tech Class)
Patent Bias(#)
10000
5000
0
ID IL DE NH MO UT SC IN OR WV CO PR WY VT AK LA SD AR MT RI NE AL KS MS NC NM ME FL KY AZ TN PA WI GA IA DC OH MA MD VA NV WA NJ CT TX MI OK MN NY CA
Patent Bias(#)/Firm(#)
25
0
−25
−50
−75
ID WV NH DE SC IL UT MO IN OR CO PR WY VT SD RI NC FL PA AL NE AR AK MA OH LA WI GA MT KS AZ CA TN NJ ME MD IA TX KY NY CT WA MI NM MN DC VA MS NV OK
State

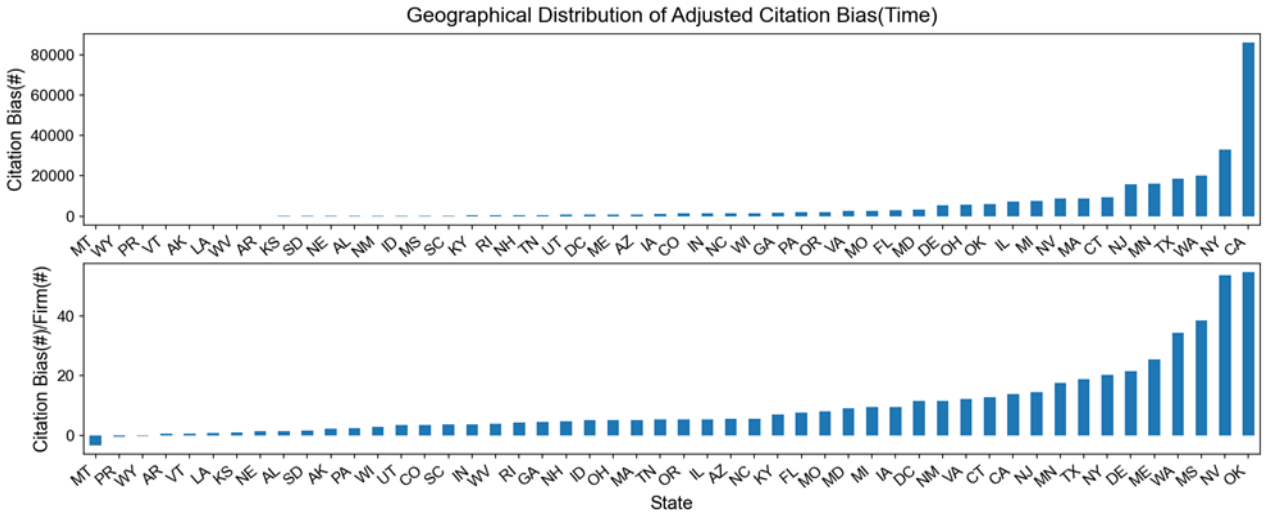
(c)

Figure 8: Firm Citation Bias (Unadjusted and Adjusted) across States

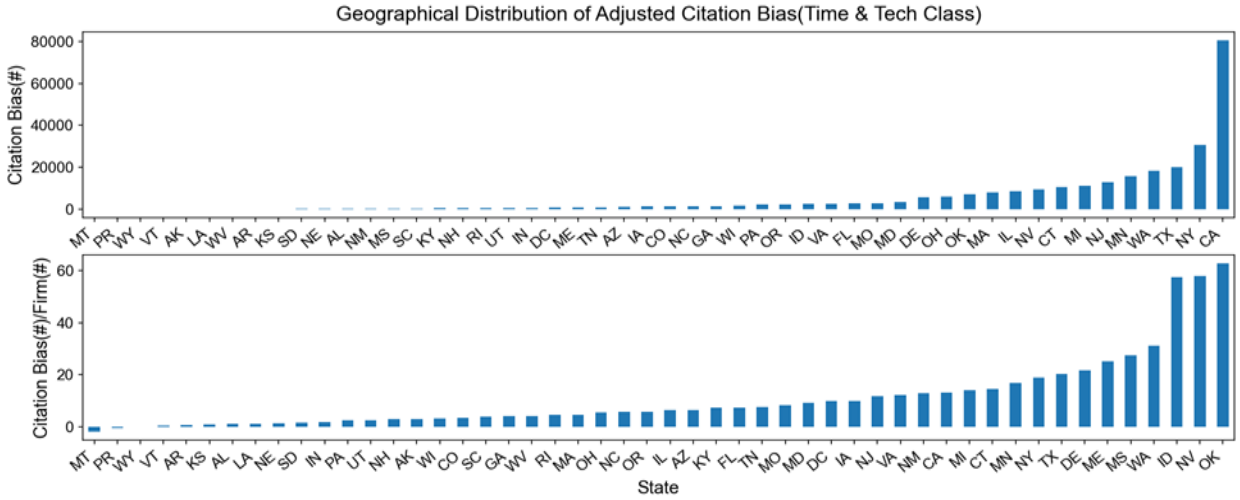
This figure presents the distribution of citation bias aggregated at the year level (upper graph in each panel) and on a mean firm-year level (lower graph) for patents granted to public firms from 1976 through 2006 in different states. To compute the unadjusted citation bias for each firm-year, we compare the number of citations to all the patents for each firm filed in each application year in our data (i.e., citations in patents granted by 2012 to applications filed by a firm in a given year and granted by 2006) and in the NBER 2006 dataset (i.e., citations in patents granted by 2006 to applications filed by a firm in a given year and granted by 2006). A firm is assigned to a particular state in a given year based on modal state of the assignee across patents granted to the firm at the time of the patent filing. The adjustments use the time fixed effect, the time and technology class fixed effect methodologies, and the citing year effect adjustment using the quasi-structural method, with details discussed in the text and Online Appendix B. Sources: NBER 2006 patent and our datasets.



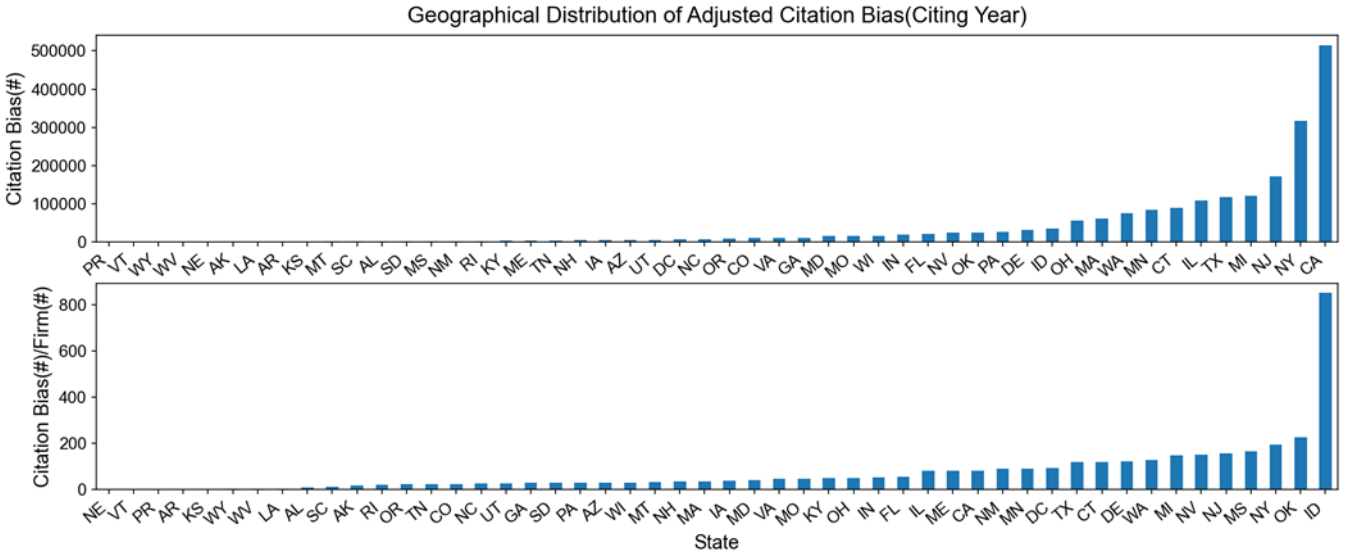
(a)


Geographical Distribution of Adjusted Citation Bias(Time)
Citation Bias(#)
Citation Bias(#)/Firm(#)
State

(b)


Geographical Distribution of Adjusted Citation Bias(Time & Tech Class)
Citation Bias(#)
Citation Bias(#)/Firm(#)
State

(c)


Geographical Distribution of Adjusted Citation Bias(Citing Year)
Citation Bias(#)
500000
400000
300000
200000
100000
0
PR VT WY WV NE AK LA AR KS MT SC AL SD MS NM RI KY ME TN NH IA AZ UT DC NC OR CO VA GA MD MO WI IN FL NV OK PA DE ID OH MA WA MN CT IL TX MI NJ NY CA
Citation Bias(#)/Firm(#)
800
600
400
200
0
NE VT PR AR KS WY WV LA AL SC AK RI OR TN CO NC UT GA SD PA AZ WI MT NH MA IA MD VA MO KY OH IN FL IL ME CA NM MN DC TX CT DE WA MI NV NJ MS NY OK ID
State

(d)

Table 1: Patent Bias and Firm Characteristics

This table presents OLS regressions relating unadjusted and adjusted patent bias at the firm level with different firm characteristics in each year between 1976 and 2006. The dependent variable is the unadjusted patent bias of a given firm in that year (columns 1-3) and the time- and technology class-adjusted patent bias of a given firm in that year (columns 4-6). The dependent variable is computed as the difference in the log of one plus the number of successful patent applications filed by a firm in a given year as of 2012 ("our data") and the log of one plus the number of successful patent applications filed by that firm in the same year as of the end of sample in the NBER 2006 dataset (unadjusted or adjusted). Control variables and their construction are described in Online Appendix D. Robust t-tests are reported in the parenthesis. Sources: NBER 2006 patent and our datasets.

	Unadjusted Patent Bias			*Adjusted Patent Bias (Time & Tech Class)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Log Size	0.0548***	0.0496***	0.0204**	0.0185***	0.0127***	0.0182**
	(13.15)	(11.60)	(2.39)	(4.02)	(2.70)	(2.38)
Log M/B	0.0464***	0.0511***	0.0305**	0.0323**	0.0378**	0.0271*
	(3.30)	(3.63)	(2.24)	(2.08)	(2.43)	(1.73)
Log RD/Sales	0.0235***	0.0333***	0.0420***	0.00756	0.0211***	0.0277**
	(3.77)	(5.14)	(4.00)	(1.10)	(2.95)	(2.29)
Log Cash/Assets	0.0205***	0.0201***	0.0156***	0.0159**	0.0157**	0.0091*
	(3.51)	(3.44)	(2.67)	(2.47)	(2.43)	(1.74)
Log Leverage	0.0722***	0.0707***	0.0461***	0.0346**	0.0331**	0.0292*
	(5.09)	(4.98)	(2.77)	(2.21)	(2.11)	(1.73)
ROA	0.0851*	0.0812*	0.125**	0.0833*	0.117**	0.173***
	(1.73)	(1.84)	(2.57)	(1.88)	(2.41)	(3.10)
Log Spread	-0.0463**	-0.0435**	-0.0850***	-0.0224	-0.0168	-0.0814***
	(-2.46)	(-2.30)	(-4.65)	(-1.08)	(-0.81)	(-3.87)
Log(Patents in State)	0.034	0.049	-1.36***	0.0598	0.0554	-1.29***
	(0.22)	(0.32)	(-3.72)	(0.35)	(0.33)	(-3.10)
Log(Patents in Technology Class)	0.0854***	0.0881***	0.00801***	0.0440**	0.0476***	0.00158*
	(5.30)	(5.49)	(3.25)	(2.47)	(2.68)	(1.69)
Observations	15331	15331	15331	15331	15331	15331
R^2	0.429	0.437	0.674	0.200	0.209	0.506
Firm Fixed Effect			Yes			Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Class Fixed Effect	Yes	Yes		Yes	Yes	
NAICS Fixed Effect	Yes			Yes		
SIC Fixed Effect		Yes			Yes	

Table 2: Citation Bias and Firm Characteristics

This table presents OLS regressions relating unadjusted and adjusted citation bias at the firm level with different firm characteristics in each year between 1976 and 2006. The dependent variable is the unadjusted citation bias of a given firm in that year (columns 1-3); the time and tech class fixed effect adjusted citation bias of a given firm in that year. The dependent variable is computed as the difference in the log of one plus the number of citations to successful patent applications of a firm in a given year as of 2012 ("our data") and the log of one plus the number of citations to successful patent applications of a firm in the same year as of the end of sample in the NBER 2006 dataset (unadjusted or adjusted). Control variables and their construction are described in Online Appendix D. Robust t-tests are reported in the parenthesis. Sources: NBER 2006 patent and our datasets.

	Unadjusted Citation Bias			*Adjusted Citation Bias (Time & Tech Class)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Log Size	0.121***	0.118***	0.0763***	0.0831***	0.0794***	0.0337***
	(15.86)	(15.13)	(4.84)	(14.18)	(13.18)	(2.63)
Log M/B	0.0920***	0.0943***	0.0618**	0.0503**	0.0541***	0.0406**
	(3.59)	(3.66)	(2.45)	(2.54)	(2.72)	(1.98)
Log RD/Sales	0.0417***	0.0479***	0.0947***	0.0308***	0.0381***	0.0661***
	(3.67)	(4.04)	(4.87)	(3.51)	(4.17)	(4.18)
Log Cash/Assets	0.0575***	0.0580***	0.0627***	0.0232***	0.0225***	0.0267***
	(5.40)	(5.44)	(5.78)	(2.82)	(2.73)	(3.03)
Log Leverage	0.0929***	0.0835***	0.108***	0.0503**	0.0431**	0.0766***
	(3.60)	(3.22)	(3.52)	(2.52)	(2.15)	(3.06)
ROA	-0.0839	-0.0623	-0.043	0.0362	0.0687	0.0767
	(-1.06)	(-0.77)	(-0.48)	(0.59)	(1.11)	(1.05)
Log Spread	-0.0783**	-0.0789**	-0.138***	-0.0545**	-0.0548**	-0.0940***
	(-2.28)	(-2.29)	(-4.09)	(-2.06)	(-2.06)	(-3.41)
Log(Citations in State)	0.501	0.448	-2.241***	0.327	0.269	-1.86***
	(1.55)	(1.56)	(-3.26)	(1.48)	(1.22)	(-3.33)
Log(Citations in Technology Class)	0.178***	0.179***	0.0215***	0.0917***	0.0935***	0.00715*
	(5.42)	(5.48)	(4.64)	(3.63)	(3.71)	(1.89)
Observations	15331	15331	15331	15331	15331	15331
R^2	0.310	0.316	0.594	0.340	0.345	0.568
Firm Fixed Effect			Yes			Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Class Fixed Effect	Yes	Yes		Yes	Yes	
NAICS Fixed Effect	Yes			Yes		
SIC Fixed Effect		Yes			Yes	

Table 3: Checklist for Analyses

This table presents a checklist for patent-based corporate finance analyses, based on the principles developed and discussed in the paper.

1. Present estimates of how patent and citation biases at firm level correlate with the key policy changes at firm, industry, or economy level being analyzed. Patent and citation bias can be computed using different versions of the patent data, using the methods discussed in this paper.

2. Present two sets of estimates related to policy changes at firm, industry or economy level being analyzed: (i) with last few years of sample being analyzed included and (ii) without the last few years in the sample being analyzed.

3. Present two sets of estimates to evaluate robustness with respect to technology class: (i) with industries that experienced a surge of patenting or in citations per patent (computers and electronics and chemicals) included in the sample and (ii) without industries that experienced a surge of patenting or in citations per patent (computers and electronics and chemicals) included in the sample.

4. Present two sets of estimates to evaluate robustness with respect to geography: (i) with states that experienced a surge of patenting or citations per patent (California and Massachusetts) included in the sample and (ii) without states that experienced a surge of patenting or citations per patent (California and Massachusetts) included in the sample.

5. Present two sets of estimates to evaluate robustness with respect to firm characteristics: (i) including firms with features that experienced a surge in patenting or citations per patent (e.g., those with a high market-to-book value and those that are large) in the sample and (ii) excluding firms with features that experienced a surge in patenting or citations per patent (e.g., those with a high market-to-book value and those that are large) in the sample.

6. Present two sets of estimates to evaluate robustness with respect to firm attrition: (i) with firms that exit over the sample period included in the sample and (ii) without firms that exit over the sample period included in the sample.

7. Present two sets of estimates that evaluate robustness with respect to limitations of the concordances used between (a) patent assignee names and (b) firm identifiers such as CUSIPs and GVKEYs: (i) including firms in the sample where the match confidence is low and (ii) excluding firms in the sample where the match confidence is low.

8. Present two sets of estimates to evaluate robustness with respect to strategic patent and citation assignment practices at the firm level: (i) including firms in the sample that might be engaging in such practices and (ii) excluding firms that might be engaging in such practices. Classification of such firms can be done using media accounts, patent assignments, and reassignment files.

Table 4: The Use of Machine Learning in Mitigating Firm-Level Patent Bias

This table presents several different machine learning models to make predictions on the number of ultimately granted patents at the firm level in each HJT class and every year. We attempt to predict the number of patent applications that will be ultimately granted to each firm for patents it files in each year and HJT technology class between 2002 and 2006. We compare the predicted values to the actual granted patents – for patents filed in these years -- recorded as of the end of 2012. We use data between 1976 and 2001 as the training data (as recorded at the end of 2006). For the purposes of the analysis, we only use patents of firms that have a Compustat identifier in the 2006 NBER database. Linear Support Vector Machine (SVM), Back Propagation Neural Network (BPNN), Decision Tree, Random Forest, Lasso Regression, and Ridge Regression are selected as the candidate models, due to their popularity in machine learning regression tasks. We evaluate the performance on three metrics: RMSE, R^2, and Correlation: For the RMSE, the smaller the metric, the better; for the R squared and Pearson product-moment correlation, the closer to one, the better. The RMSE and R squared from ML predictions are compared to those using raw benchmark (the number of ultimately granted patents at the firm level in each HJT class as of 2006), the time-adjusted benchmark, and the enhanced time- and technology-adjusted benchmark. The ML models use past patenting activity and patent categories as features. More details and other variants of the machine learning models are discussed in Online Appendix I. Sources: NBER 2006 patent and our datasets.

Model/Metric	Root Mean Squared Error(RMSE)	R Squared (R^2)	Correlation	Better Than Raw Benchmark?	Better Than Time-Adjusted BM?	Better Than Enhanced Time- and Technology-Adjusted BM?
Raw Benchmark	113.25	0.43	0.73	--	NO	NO
Time-Adjusted Benchmark	96.92	0.59	0.84	YES	--	NO
Enhanced Time- and Technology-Adjusted Benchmark	91.47	0.63	0.84	YES	YES	--
Linear SVR	66.22	0.81	0.90	YES	YES	YES
BP Neural Network	76.76	0.74	0.87	YES	YES	YES
Decision Tree	80.56	0.71	0.86	YES	YES	YES
Random Forest	74.57	0.75	0.88	YES	YES	YES
Lasso Regression	69.44	0.79	0.89	YES	YES	YES
Ridge Regression	70.68	0.78	0.88	YES	YES	YES

Table 5: The Use of Machine Learning in Mitigating Firm-Level Citation Bias

This table presents several different machine learning models to make predictions on the number of citations ten years after ultimately granted patents at the firm level every year. The models are trained with the citation counts for patents applied for between 1976 and 1992 in the NBER 2006 data set. We predict the number of citations in the ten years after issuance for each patent granted to a selected firm in the years from 1993 to 2002. We compare the predicted values to the actual citations -- for patents filed in these years -- recorded as of the end of 2012. For the purposes of the analysis, we only use patents of firms that have a Compustat identifier in the 2006 NBER database. Linear Support Vector Machine (SVM), Back Propagation Neural Network (BPNN), Decision Tree, Random Forest, Lasso Regression, and Ridge Regression are selected as the candidate models, due to their popularity in machine learning regression tasks. We evaluate the performance on three metrics: RMSE, R^2, and Correlation: For the RMSE, the smaller the metric, the better; for the R squared and Pearson product-moment correlation, the closer to one, the better. The RMSE and R squared from ML predictions are compared to those using raw benchmark (the number of citations as of 2006), the time-adjusted benchmark, the enhanced time-and technology-adjusted benchmark, and the quasi-structural benchmark. The ML models use past patenting activity and patent categories as features. More details and other variants of the machine learning models are discussed in Online Appendix I. Sources: NBER 2006 patent and our datasets.

Model/Metric	Root Mean Squared Error(RMSE)	R Squared (R^2)	Correlation	Better Than Raw Benchmark?	Better Than Time-Adjusted BM?	Better Than Enhanced Time- and Technology-Adjusted BM?	Better Than Quasi-Structural BM?
Raw Benchmark	1266.43	-0.46	0.90	--	NO	NO	MIXED
Time-Adjusted Benchmark	818.14	0.69	0.96	YES	--	NO	MIXED
Enhanced Time- and Technology-Adjusted Benchmark	814.87	0.69	0.96	YES	YES	--	MIXED
Quasi-Structural Benchmark	1852.32	0.75	0.95	MIXED	MIXED	MIXED	--
Linear SVR	733.60	0.77	0.97	YES	YES	YES	YES
BP Neural Network	612.46	0.90	0.96	YES	YES	YES	YES
Decision Tree	752.21	0.85	0.93	YES	YES	YES	YES
Random Forest	697.06	0.85	0.95	YES	YES	YES	YES
Lasso Regression	656.77	0.85	0.96	YES	YES	YES	YES
Ridge Regression	656.80	0.85	0.96	YES	YES	YES	YES

Appendix: Papers using Patent Count and/or Citation Data in Major Finance Journals, 2005-2020

Journal of Finance

Atanassov, Julian, 2013, "Do Hostile Takeovers Stifle Innovation? Evidence from Antitakeover Legislation and Corporate Patenting," *Journal of Finance* 68, 1097-1131.

Bartram, Söhnke M., Gregory Brown, and René M. Stulz, 2012, "Why Are U.S. Stocks More Volatile?," *Journal of Finance* 67, 1329-1370.

Bena, Jan, and Kai Li, 2014, "Corporate Innovations and Mergers and Acquisitions," *Journal of Finance* 69, 1923-1960.

Bernstein, Shai, 2015, "Does Going Public Affect Innovation?," *Journal of Finance* 70, 1365-1403.

Bernstein, Shai, Xavier Giroud, and Richard R. Townsend, 2016, "The Impact of Venture Capital Monitoring," *Journal of Finance* 71, 1591–1622.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist, 2020, "What Is a Patent Worth? Evidence from the U.S. Patent 'Lottery,'" *Journal of Finance*, 75, 639-682.

Gompers, Paul, Josh Lerner, and David Scharfstein, 2005, "Entrepreneurial Spawning: Public Corporations and the Genesis of New Ventures, 1986 to 1999," *Journal of Finance* 60, 577-614.

Haslem, Bruce, 2005, "Managerial Opportunism during Corporate Litigation," *Journal of Finance* 60, 2013-2041.

Hirshleifer, David, Angie Low, and Siew Hong Teoh, 2012, "Are Overconfident CEOs Better Innovators?," *Journal of Finance* 67, 1457-1498.

Hoberg, Gerard, and Gordon Phillips, 2010, "Real and Financial Industry Booms and Busts," *Journal of Finance* 65, 45-86.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala, 2014, "Product Market Threats, Payouts, and Financial Flexibility," *Journal of Finance* 69, 293-324.

Hombert, Johan and Adrien Matray, 2018, "Can Innovation Help U.S. Manufacturing Firms Escape Import Competition from China?," *Journal of Finance*, 73, 2003-2039.

Hsu, David H., 2004, "What Do Entrepreneurs Pay for Venture Capital Affiliation?," *Journal of Finance* 59, 1805-1844.

Kaplan, Steven N., Berk A. Sensoy, and Per Strömberg, 2009, "Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies," *Journal of Finance* 64, 75-115.

Karpoff, Jonathan M. and Michael D. Wittry, 2018, "Institutional and Legal Context in Natural Experiments: The Case of State Antitakeover Laws," *Journal of Finance*, 73, 657-714.

Kumar, Praveen, and Dongmei Li, 2016, "Capital Investment, Innovative Capacity, and Stock Returns," *Journal of Finance* 71, 2059–2094.

Kung, Howard, and Lukas Schmid, 2015, "Innovation, Growth, and Asset Prices," *Journal of Finance* 70, 1001-1037.

Lerner, Josh, Morten Sorensen, and Per Strömberg, 2011, "Private Equity and Long-Run Investment: The Case of Innovation," *Journal of Finance* 66, 445-477.

Journal of Financial Economics

Acharya, Viral, and Zhaoxia Xu, 2017, "Financial Dependence and Innovation: The Case of Public versus Private Firms," *Journal of Financial Economics* 124, 223-243.

Ali, Usman, and David Hirshleifer, 2020, "Shared Analyst Coverage: Unifying Momentum Spillover Effects," *Journal of Financial Economics*, 136, 649-675.

Amore, Mario D., Cédric Schneider, and Alminas Žaldokas, 2013, "Credit Supply and Corporate Innovation," *Journal of Financial Economics* 109, 835-855.

Appel, Ian, Joan Farre-Mensa, and Elena Simintzi, 2019, "Patent Trolls and Startup Employment," *Journal of Financial Economics* 133, 708-725.

Balsmeier, Benjamin, Lee Fleming, and Gustavo Manso, 2017, "Independent Boards and Innovation," *Journal of Financial Economics* 123, 536-557.

Baranchuk, Nina, Robert Kieschnick, and Rabih Moussawi, 2014, "Motivating Innovation in Newly Public Firms," *Journal of Financial Economics* 111, 578-588.

Bena, Jan, Miguel A. Ferreira, Pedro Matos, and Pedro Pires, 2017, "Are Foreign Investors Locusts? The Long-Term Effects of Foreign Institutional Ownership," *Journal of Financial Economics* 126, 122-146.

Benson, David, and Rosemarie H. Ziedonis, 2010, "Corporate venture capital and the returns to acquiring portfolio companies," *Journal of Financial Economics* 98, 478-499.

Bernile, Gennaro, Vineet Bhagwat, and Scott Yonker, 2018, "Board Diversity, Firm Risk, and Corporate Policies", *Journal of Financial Economics*, 127, 588-612.

Bernstein, Shai, Abhishek Dev, and Josh Lerner, 2020, "The Creation and Evolution of Entrepreneurial Public Markets," *Journal of Financial Economics*, 136, 307-329.

Blanco, Iván, and David Wehrheim, 2017, "The Bright Side of Financial Derivatives: Options Trading and Firm Innovation," *Journal of Financial Economics* 125, 99-119.

Brav, Alon, Wei Jiang, Song Ma, and Xuan Tian, 2018, "How Does Hedge Fund Activism Reshape Corporate Innovation?," *Journal of Financial Economics*, 130, 237-264.

Campello, Murillo, and Janet Gao, 2017, "Customer Concentration and Loan Contract Terms," *Journal of Financial Economics* 123, 108-136.

Chang, Xin, Yangyang Chen, Sarah Qian Wang, Kuo Zhang, and Wenrui Zhang, "Credit Default Swaps and Corporate Innovation," *Journal of Financial Economics*, 134, 474-500.

Chang, Xin, Kangkang Fu, Angie Low, and Wenrui Zhang, 2015, "Non-Executive Employee Stock Options and Corporate Innovation," *Journal of Financial Economics* 115, 168-188.

Chava, Sudheer, Alexander Oettl, Ajay Subramanian, and Krishnamurthy V. Subramanian, 2013, "Banking Deregulation and Innovation," *Journal of Financial Economics* 109, 759-774.

Cornaggia, Jess, Yifei Mao, Xuan Tian, and Brian Wolfe, 2015, "Does Banking Competition Affect Innovation?," *Journal of Financial Economics* 115, 189-209.

Cornaggia, Jess and Jay Yin Li, "The Value of Access to Finance: Evidence from M&As," *Journal of Financial Economics,* 131, 232-250.

Cremers, K.J. Martijn, Lubomir P. Litov, and Simone M. Sepe, 2017 "Staggered Boards and Long-Term Firm Value, Revisited," *Journal of Financial Economics*, 126, 422-444.

Croce, M. M., Thien T. Nguyen, S. Raymond, and L. Schmid, "Government Debt and the Returns to Innovation," *Journal of Financial Economics*, 132, 205-225.

Custódio, Cláudia, and Daniel Metzger, 2014, "Financial Expert CEOs: CEO's Work Experience and Firm's Financial Policies," *Journal of Financial Economics* 114, 125-154.

Dyreng, Scott D., Bradley P. Lindsey, and Jacob R. Thornock, 2013, "Exploring The Role Delaware Plays as a Domestic Tax Haven," *Journal of Financial Economics* 108, 751-772.

Faleye, Olubunmi, Rani Hoitash, and Udi Hoitash, 2011, "The Costs of Intense Board Monitoring," *Journal of Financial Economics* 101, 160-181.

Frydman, Carola, and Dimitris Papanikolaou, 2018, "In Search of Ideas: Technological Innovation and Executive Pay Inequality," *Journal of Financial Economics*, 130, 1-24.

Gomes-Casseres, Benjamin, John Hagedoorn, and Adam B. Jaffe, 2006, "Do Alliances Promote Knowledge Flows?," *Journal of Financial Economics* 80, 5-33.

González-Uribe, Juanita, 2020, "Exchanges of Innovation Resources inside Venture Capital Portfolios," *Journal of Financial Economics*, 135, 144-168.

Guo, Bing, David Pérez-Castrillo, and Anna Toldrà-Simats, 2019, "Firms' Innovation Strategy under the Shadow of Analyst Coverage," *Journal of Financial Economics*, 131, 456-483.

Gu, Tiantian, 2017, "U.S. Multinationals and Cash Holdings," *Journal of Financial Economics* 125, 344-368.

He, Jie, and Xuan Tian, 2013, "The Dark Side of Analyst Coverage: The Case of Innovation," *Journal of Financial Economics* 109, 856-878.

Higgins, Matthew J., and Daniel Rodriguez, 2006, "The Outsourcing of R&D through Acquisitions in the Pharmaceutical Industry," *Journal of Financial Economics* 80, 351-383.

Hochberg, Yael V., Carlos J. Serrano, and Rosemarie H. Ziedonis, 2018, "Patent Collateral, Investor Commitment, and the Market for Venture Lending," *Journal of Financial Economics*, 130, 74-94.

Hirshleifer, David, Po-Hsuan Hsu, and Dongmei Li, 2013, "Innovative Efficiency and Stock Returns," *Journal of Financial Economics* 107, 632-654.

Hsu, Po-Hsuan, 2009, "Technological Innovations and Aggregate Risk Premiums," *Journal of Financial Economics* 94, 264-279.

Hsu, Po-Hsuan, Xuan Tian, and Yan Xu, 2014, "Financial Development and Innovation: Cross-Country Evidence," *Journal of Financial Economics* 112, 116-135.

Humphery-Jenner, Mark, Ling L. Lisic, Vikram Nanda, and Sabatino D. Silveri, 2016, "Executive Overconfidence and Compensation Structure," *Journal of Financial Economics* 119, 533-558.

Islam, Emdad, and Jason Zein, 2020, "Inventor CEOs," *Journal of Financial Economics*, 135, 505-527.

Jones, Christopher S., and Selale Tuzel, 2013, "Inventory Investment and the Cost Of Capital," *Journal of Financial Economics* 107, 557-579.

Kim, Hyunseob, 2020, "How Does Labor Market Size Affect Firm Capital Structure? Evidence from Large Plant Openings," *Journal of Financial Economics*, 138, 277-294.

Kwon, Sungjoung, Michelle Lowry, and Yiming Qian, 2020, "Mutual Fund Investments in Private Firms," *Journal of Financial Economics*, 136, 407-443.

Lee, Charles M. C., Stephen Teng Sun, Rongfei Wang, and Ran Zhang, "Technological Links and Predictable Returns," *Journal of Financial Economics,* 132, 76-96.

Lerner, Josh, 2006, "The New New Financial Thing: The Origins of Financial Innovations," *Journal of Financial Economics* 79, 223-255.

Lerner, Josh, Antoinette Schoar, Stanislav Sokolinski, and Karen Wilson, 2018, "The Globalization of Angel Investments: Evidence across Countries," *Journal of Financial Economics* 127, 1-20.

Mann, William, 2018, "Creditor Rights and Innovation: Evidence from Patent Collateral," *Journal of Financial Economics,* 130, 25-47.

Mukherjee, Abhiroop, Manpreet Singh, and Alminas Žaldokas, 2017, "Do Corporate Taxes Hinder Innovation?," *Journal of Financial Economics* 124, 195-221.

Na, Ke, 2020, "CEOs' Outside Opportunities and Relative Performance Evaluation: Evidence from a Natural Experiment," *Journal of Financial Economics,* 137, 679-700.

Nanda, Ramana, and Matthew Rhodes-Kropf, 2013, "Investment Cycles and Startup Innovation," *Journal of Financial Economics* 110, 403-418.

Nanda, Ramana, and Tom Nicholas, 2014, "Did Bank Distress Stifle Innovation during the Great Depression?," *Journal of Financial Economics* 114, 273-292.

Ozmel, Umit, David T. Robinson, and Toby E. Stuart, 2013, "Strategic Alliances, Venture Capital, and Exit Decisions in Early Stage High-Tech Firms," *Journal of Financial Economics* 107, 655-670.

Qiu, Jiaping, and Chi Wan, 2015, "Technology Spillovers and Corporate Cash Holdings," *Journal of Financial Economics* 115, 558-573.

Segal, Gill, Ivan Shaliastovich, and Amir Yaron, 2015, "Good and Bad Uncertainty: Macroeconomic and Financial Market Implications," *Journal of Financial Economics* 117, 369-397.

Seru, Amit, 2014, "Firm Boundaries Matter: Evidence from Conglomerates and R&D Activity," *Journal of Financial Economics* 111, 381-405.

Sunder, Jayanthi, Shyam V. Sunder, and Jingjing Zhang, 2017, "Pilot CEOs and Corporate Innovation, *Journal of Financial Economics* 123, 209-224.

Review of Financial Studies

Acharya, Viral V., and Krishnamurthy V. Subramanian, 2009, "Bankruptcy Codes and Innovation," *Review of Financial Studies* 22, 4949-4988.

Acharya, Viral V., Ramin P. Baghai, and Krishnamurthy V. Subramanian, 2014, "Wrongful Discharge Laws and Innovation," *Review of Financial Studies* 27, 301-346.

Agarwal, Vikas, Rahul Vashishtha, and Mohan Venkatachalam, 2018, "Mutual Fund Transparency and Corporate Myopia," *Review of Financial Studies*, 31, 1966–2003
.
Bai, John (Jianqiu), Douglas Fairhurst, and Matthew Serfling, 2020, "Employment Protection, Investment, and Firm Growth," *Review of Financial Studies*, 33, 644–688.

Ball, Eric, Hsin Hui Chiu, and Richard Smith, 2011, "Can VCs Time the Market? An Analysis of Exit Choice for Venture-backed Firms," *Review of Financial Studies* 24, 3015-3138.

Bircan, Çağatay, and Ralph De Haas, 2020, "The Limits of Lending? Banks and Technology Adoption across Russia," *Review of Financial Studies*, 33, 536–609.

Celikyurt, Ugur, Merih Sevilir, and Anil Shivdasani, 2014, "Venture Capitalists on Boards of Mature Public Firms," *Review of Financial Studies* 27, 56-101.

Chari, Anusha, Paige P. Ouimet, and Linda L. Tesar, 2010, "The Value of Control in Emerging Markets," *Review of Financial Studies* 23, 1741-1770.

Chemmanur, Thomas J., Elena Loutskina, and Xuan Tian, 2014, "Corporate Venture Capital, Value Creation, and Innovation," *Review of Financial Studies* 27, 2434-73.

Chen, Mark A., Qinxi Wu, and Baozhong Yang, 2019, "How Valuable Is FinTech Innovation?," *Review of Financial Studies*, 32, 2062–2106.

Cohen, Lauren, Karl Diether, and Christopher Malloy, 2013, "Misvaluing Innovation," *Review of Financial Studies* 26, 635-666.

Dasgupta, Sudipto, and Alminas Žaldokas, 2019, "Anticollusion Enforcement: Justice for Consumers and Equity for Firms," *Review of Financial Studies*, 32, 2587–2624.

Dass, Nishant, Omesh Kini, Vikram Nanda, Bunyamin Onal, and Jun Wang, 2014, "Board Expertise: Do Directors from Related Industries Help Bridge the Information Gap?," *Review of Financial Studies* 27, 1533-92.

Duval, Romain, Gee Hee Hong, and Yannick Timmer, 2020, "Financial Frictions and the Great Productivity Slowdown," *Review of Financial Studies*, 33, 475–503.

Fang, Lily H., Josh Lerner, and Chaopeng Wu, 2017, "Intellectual Property Rights Protection, Ownership, and Innovation: Evidence from China," *Review of Financial Studies* 30, 2446–77.

Frésard, Laurent, Ulrich Hege, and Gordon Phillips. 2017, "Extending Industry Specialization through Cross-Border Acquisitions," *Review of Financial Studies* 30, 1539–82.

Frésard, Laurent, Gerard Hoberg, and Gordon M. Phillips, 2020, "Innovation Activities and Integration through Vertical Acquisitions," *Review of Financial Studies*, 33, 2937–2976.

Gan, Jie, Yan Guo, and Chenggang Xu, 2018, "Decentralized Privatization and Change of Control Rights in China," *Review of Financial Studies*, 31, 3854-3894.

Grieser, William, and Zack Liu, 2019, "Corporate Investment and Innovation in the Presence of Competitor Constraints," *Review of Financial Studies*, 32, 4271–4303.

He, Jie, and Jiekun Huang, 2017, "Product Market Competition in a World of Cross-Ownership: Evidence from Institutional Blockholdings," *Review of Financial Studies* 30, 2674–2718.

Heath, Davidson, and Christopher Mace, 2020, "The Strategic Effects of Trademark Protection," *Review of Financial Studies*, 33, 1848–1877.

Hirshleifer, David, Po-Hsuan Hsu, and Dongmei Li, 2018, "Innovative Originality, Profitability, and Stock Returns," *Review of Financial Studies*, 31, 2553-2605.

Hombert, Johan, and Adrien Matray, 2017, "The Real Effects of Lending Relationships: Evidence from Banking Deregulation and Innovation," *Review of Financial Studies* 30, 2413-45.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020, "Replicating Anomalies," *Review of Financial Studies*, 33, 2019–2133.

Irvine, Paul J., and Jeffrey Pontiff, 2009, "Idiosyncratic Return Volatility, Cash Flows, and Product Market Competition," *Review of Financial Studies* 22, 1149-1177.

Kerr, William R., Josh Lerner, and Antoinette Schoar, 2014, "The Consequences of Entrepreneurial Finance: Evidence from Angel Financings," *Review of Financial Studies* 27, 20-55.

Ma, Song, 2020, "The Life Cycle of Corporate Venture Capital," *Review of Financial Studies*, 33, 358–394.

Ouimet, Paige P., 2013, "What Motivates Minority Acquisitions? The Trade-Offs between a Partial Equity Stake and Complete Integration," *Review of Financial Studies* 26, 1021-1047.

Phillips, Gordon M., and Giorgo Sertsios, 2017, "Financing and New Product Decisions of Private and Publicly Traded Firms," *Review of Financial Studies* 30, 1744–1789.

Tian, Xuan, and Tracy Yue Wang, 2011, "Tolerance for Failure and Corporate Innovation," *Review of Financial Studies* 27, 211-255.