Screening Property Rights for Innovation[*]

William Matcham[§] Mark Schankerman[§]

October 5, 2023

Abstract

We develop a dynamic structural model of patent screening incorporating incentives, intrinsic motivation, and multi-round negotiation. We estimate the model using detailed data on examiner decisions and employ natural language processing to create a new measure of patent distance that enables us to study strategic decisions by applicants and examiners. The estimated parameters and counterfactual analysis imply three main findings. First, patent screening is moderately effective, *given* the existing standards for patentability. Second, examiners exhibit substantial intrinsic motivation that significantly improves the effectiveness, and reduces the net social costs, of screening. Third, limiting the number of negotiation rounds strongly increases the speed and quality of screening. We quantify the net social costs of patent screening and find that the annual social cost of the existing system is $25.5bn, equivalent to 6.5% of U.S. R&D performed by the private sector.

Keywords: Patents, innovation, incentives, screening, intrinsic motivation
JEL Classification: D73, L32, O31, O34, O38

[*] We acknowledge Jakub Drabik for his excellent research assistance and contributions, especially on the distance metric. We thank Dietmar Harhoff, Marco Ottaviani, Florian Schuett, and seminar and conference participants at Tel Aviv University, Humboldt University of Berlin, Ben-Gurion University, CEMFI, University of Bologna, Royal Holloway, and Boston University for comments and suggestions on earlier drafts of the paper. We also thank Martin Rater, Daniel Ryman and Andrew Toole in the U.S. Patent and Trademark Office for assistance in obtaining part of the data, and Janet Freilich and Michael Meurer for helpful discussions on procedural aspects of patent prosecution. This project was partly financed by a research grant from the Suntory-Toyota Centres for Economic and Related Disciplines at the London School of Economics. William Matcham also thanks the Edison Fellowship Program at George Mason University for financial assistance and the opportunity to discuss the project with legal scholars.

[§] Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Emails: w.o.matcham@lse.ac.uk, m.schankerman@lse.ac.uk

1 Introduction

Public institutions play a central role in promoting innovation. The two most important channels are government support for public and private research, both in the form of direct funding and indirect fiscal subsidies, and the allocation of property rights, in the form of patents, to enhance innovation incentives for private sector R&D. To give a sense of the scale of investment, in 2015 the U.S. federal government financed 54.3% of overall R&D expenditures, or $151.5 billion (2023 U.S.D.), and 34.1% of university research. At the same time, the U.S. Patent and Trademark Office (hereafter, *Patent Office*) issued nearly 400,000 new patents. These property rights promote innovation by increasing the private returns to R&D, facilitating access to capital markets, and underpinning the market for technology, especially for small, high-technology firms (Hall and Lerner, 2010; Galasso and Schankerman, 2018). Moreover, the aggregate economic impact of these investments and property rights for innovation is magnified by the extensive knowledge spillovers they generate (Bloom, Schankerman, and Van Reenen, 2013).

Despite their evident importance, little is known about whether innovation-supporting public institutions allocate resources efficiently and how organizational changes affect agency performance. The aim of this paper, as part of a broader research program, is to show how structural models can be used to study and improve the efficiency of resource allocation by innovation-related public agencies. We study this topic in the context of the U.S. patent system, focusing on the quality of screening—the allocation of property rights for innovation—by the Patent Office.

We develop a dynamic structural model of the patent screening process, which incorporates incentives, intrinsic motivation, and the actual structure of multi-round negotiation in the current system. We estimate the model using novel negotiation-round-level data on examiner decisions and text data from 20 million patent claims. From the claim text data, we use modern natural language processing (NLP) methods to develop a new measure of distance between patents, a key ingredient for characterizing strategic decisions by patent applicants and examiners. We conduct counterfactual analyses of how reforms to incentives, fees, and the structure of negotiations affect the quality and speed of patent screening, and we develop an approach to quantify these impacts and thus construct a "pseudo-welfare" measure of the quality of patent screening.

The effectiveness of patent screening and its implications for the quality of patents is a hotly debated policy issue. Academic scholars and policymakers have argued that patent rights have increasingly become an impediment to innovation rather than an incentive. These concerns have been prominently voiced in public debates (The Economist, 2015; Federal Trade Commission, 2011), recent U.S. Supreme Court decisions (eBay Inc. v. MercExchange L.L.C., 547 U.S. 338, 2006), and the major statutory reform of the patent system, the Leahy-Smith America Invents Act of 2011.

Critics of the patent system claim that the problems arise in large part from ineffective patent office screening, where patents are granted to inventions that do not represent a substantial inventive step – especially in emerging technology areas such as business methods and software (Jaffe and Lerner, 2004). The issue is important because granting "excessive" patent rights imposes static and dynamic social costs: higher prices and deadweight loss on patented goods, greater enforcement (litigation) costs, and higher transaction costs of R&D and the potential for retarding cumulative innovation (Galasso and Schankerman, 2015).

The patent prosecution process is an advantageous context to study the effects of incentives and motivation on screening for two primary reasons. First, the patent application process has a clear and well-documented structure that can be modeled. The multi-round negotiation between the applicant and examiner fits naturally into a dynamic game, which forms the basis of our model. The model involves an applicant who "pads" their patent application, attempting to extract more property rights than their invention truly entails. The examiner's role is to grant or reject the application based on the existing judicial interpretation of statutory criteria as applied to each claim in the patent application.[1]

The fundamental trade-off for the applicant when choosing the level of padding is between the benefits of increased patent scope and the costs of engaging in a lengthy and costly negotiation with the examiner. The trade-off for the patent examiner for each specific application is between the incentives to grant patents quickly and the intrinsic utility cost of awarding an inappropriate degree of "patent scope" – i.e., granting only patent claims (after narrowing) that satisfy the patentability criteria.[2] The patent examiner searches prior art to estimate the appropriate scope of patent protection for the invention, but this estimate contains error. Allowing for examiner error is important because it implies that negotiation between the applicant and examiner, while costly, may not always be socially wasteful.

The second advantage of the patent context is the quality of data. The Patent Office collects

[1] The two key criteria are the novelty and size of the inventive step – or conversely, how close the invention is to existing patents (non-obviousness) – and the requirement that the patent application clearly specify the relationship between the invention and the scope of rights claimed by the applicant (indefiniteness). In the model, we focus on the first criterion, but discuss an extension that allows for the second. For details, see Section 5 and Appendix E.

[2] For a discussion of the economics and legal doctrines of patent scope, see Merges and Nelson (1990).

detailed and extensive data on all *applications*, not just granted patents. For this paper, we constructed a dataset covering around 55 million patent application decisions across 20 million patent claims between 2010–2015. For applications, we observe the examiner's decisions on each patent claim over all rounds of the negotiation. We also use methods from NLP to create distances between patent claims based on claim text data. Together with the characteristics of examiners and applicants, we use these data to estimate our model of the patent application process. Ours is the first paper to incorporate intrinsic motivation into an estimated dynamic model of screening in a public agency.

Our estimates imply several key empirical findings. First, intrinsic motivation plays a significant role in contributing to the accuracy of patent screening. Junior examiners are more motivated than seniors on average, but both groups display substantial heterogeneity. Further, using the estimated parameters, counterfactual analysis shows that turning off intrinsic motivation increases the frequency of examiners granting invalid patents four–fold. This finding highlights the importance of designing human resource policies that effectively select examiners with high intrinsic motivation and ensure examiners sustain this motivation over their entire careers.

Second, we find that innovators substantially pad their patent applications, claiming (typically) greater property rights than are warranted by the true "inventive step" of their innovation. Moreover, there is a large degree of heterogeneity in the extent of padding across patent applications. This result highlights the importance of effective screening. An essential feature of our model is that the extent of padding is endogenous and thus is affected by various counterfactual policy reforms, which we detail later. We estimate the average level of padding at about 8%, rising to 10% when we weight by the value of the patent. This exaggerated scope of the patent applications is reflected in the fact that more than 80% of *claims* start below the distance threshold for patentability—as measured by the minimum required distance to claims in prior patents—and thus should be rejected.

However, the multi-round screening process is relatively effective at narrowing the scope of patent rights sought and, in so doing, reducing the number of invalid claims to about 7% among granted claims, but still, nearly one in five granted patents contains at least one claim that does not meet the threshold. One implication of this finding is that the proportion of patent applications that are granted—a commonly used indicator of the effectiveness of screening—is a misleading measure because it does not capture the extent to which granted property rights are narrowed during the screening process.

We use the estimated parameters from the baseline model to evaluate counterfactual reforms involving changes in fees for the patent applicant, the structure of the negotiation process (e.g., limiting the number of rounds allowed), and the degree of intrinsic motivation of patent examiners. We quantify the effects of counterfactual reforms along three distinct dimensions. The first two relate to the accuracy of screening, meaning the degree of alignment between the scope of property rights granted and the scope justified by the invention. We assess accuracy in terms of granting claims that are not justified (false grants, or "type 1" error) and not granting claims that should be (false rejections or "type 2" error). Each of these errors carries its own social costs and benefits. Incorrect grants impose ex post welfare costs (deadweight loss) from higher prices and litigation costs associated with enforcing these patents, but at the same timem may raise innovation incentives. False rejections dilute ex ante innovation incentives and discourage the development of new inventions that would contribute positive social value, but at the same time they reduce ex post deadweight loss. The last dimension is the speed of patent examination, measured by the number of negotiation rounds in equilibrium. We develop a method to quantify these impacts in terms of the associated net social costs (social costs net of social benefits) and thus construct a "pseudo-welfare" measure of the quality of patent screening.

We estimate the total net social cost of patent screening at $25.5bn per annual cohort of applications. This figure represents 6.5% of total R&D performed by business enterprises in the United States. Social costs include the administrative cost to the patent office, the transaction cost for applicants, the ex post cost and ex ante benefit of granting patents that do not meet the standard, and the ex ante (incentive) costs and ex post benefit of not granting patents that do meet the standard.

The counterfactual analysis highlights two key conclusions. First, restrictions on the number of allowable rounds of negotiation (currently absent in the U.S. patent system) significantly reduce the net social costs of screening, with reductions up to 45% depending on the severity of the restrictions. We show that these outcomes can be replicated through an equivalent fee per round for the applicant, but the required fees are too high to be politically feasible. Second, given the high levels of intrinsic motivation we estimate, extrinsic incentives are largely ineffective, leading to almost no change in net social costs. However, extrinsic incentives do affect outcomes in a scenario with low intrinsic motivation. For example, removing existing examiner rewards after the first examination round leads examiners to increase first-round grants by nearly 10%, even though applicants double the extent of their padding. This is associated with a lower net social cost of screening, so in this sense, extrinsic incentives are counterproductive.

The paper is organized as follows. Section 2 briefly summarizes the related literature. Section 3 describes features of the patent examination process which guide our modeling choices. Section

4 describes the datasets and summarizes key descriptive features. The structural model is presented in Section 5. Section 6 describes our estimation methods. Section 7 presents the empirical estimates. Section 8 analyzes the impact of counterfactual reforms on the accuracy and speed of patent screening, and Section 9 describes our quantification of the net social costs and benefits associated with these counterfactual reforms. Section 10 concludes.

2 Related Literature

Intrinsic Motivation in Public Agencies

We contribute to the literature that studies how intrinsic motivation affects the optimal design of incentives in mission-oriented agencies. Leading theoretical articles include Benabou and Tirole (2003; 2006), Besley and Ghatak (2005), and Prendergast (2007). Benabou and Tirole (2003; 2006) show conditions under which extrinsic rewards may crowd out intrinsic motivation. Particularly relevant to our paper, Besley and Ghatak (2005) emphasize how intrinsic motivation—which they define as the alignment between worker and agency objectives—induces welfare-improving sorting of workers across entities with different goals. They also show how intrinsic motivation affects the optimal design of incentives and authority.

There are also empirical studies using field experiments to study intrinsic motivation and public agency performance, which rely on various proxies for motivation. A leading example is Ashraf, Bandiera, and Jack (2014), who evaluate the effect of two types of extrinsic rewards (financial and non-financial) on agents' performance in a public health organization in Zambia. They find that extrinsic rewards and intrinsic motivation are complementary: both types of extrinsic rewards improve performance, but their effects are more significant for "pro-socially" motivated agents. In a related paper, Ashraf, Bandiera, Davenport, and Lee (2020) study whether career benefits attract talent at the expense of "pro-social" motivation. Except for low skill levels, there is no apparent trade-off, and pro-social motivation is associated with more effort and better performance.

Despite their interesting findings, these empirical studies cannot be used for proper counterfactual policy analysis, for which structural models are more appropriate. Our paper is the first to incorporate heterogeneous intrinsic motivation in a structural model of a public agency.[3] In doing this, we follow Besley and Ghatak's definition of intrinsic motivation – alignment of workers' objectives and the public agency mission. In our context, the Patent Office's mission is to award

[3] Egan, Matvos, and Seru (2018) develop a structural model of consumer arbitration in which arbitrators differ in their idiosyncratic degrees of "slant" (or bias), which can be interpreted as a form of intrinsic motivation.

inventors property rights over their invention, consistent with statutory and judicial prescriptions. We model intrinsic motivation as an inherent disutility that examiners incur if they grant more intellectual property rights than they believe the inventor deserves, based on the information the examiners have. In this setting, we show that patent examiners sometimes award patents to applications they believe are invalid due to strategic considerations and the extrinsic pay scheme they face.

Finally, there are recent papers that study how screening mechanisms affect the performance of public agencies. In a significant theoretical contribution, Adda and Ottaviani (2023) develop a model of nonmarket allocation of resources, including but not limited to the award of grants to research projects. The model incorporates endogenous self-selection of projects and noisy information on project quality for purposes of evaluation. The authors study how the design of allocation rules affects equilibrium applications in different fields, and how the informational noise affects the optimal design. In two empirical papers, Li and Agha (2015) and Li (2017) analyze the allocation of research grants at the National Institutes of Health (NIH). They show that peer review increases the effectiveness of grants in terms of post-grant citations but also that, while more experienced peer reviewers are better informed, they are also more biased about the quality of projects in their area of expertise, implying a trade-off. Azoulay, Graff Zivin, Li, and Sampat (2018) study the economic impact of these NIH grants, linking screening outcomes to publication citations and other innovation outcomes. Our contribution is to quantify some of the forces these papers identify—in particular, incentives, motivation, and seniority—and evaluate the equilibrium effects of various counterfactual reforms in the patent context.

Patents and Innovation

We also contribute to the limited empirical literature on patent screening. In a first paper on the topic, Cockburn, Kortum, and Stern (2003) show that patent examiner characteristics affect the "quality" of issued patents, measured by subsequent citations and the frequency of litigation (see Lemley and Sampat (2012) for additional evidence). Frakes and Wasserman (2017) exploit detailed data on promotions of patent examiners (which are accompanied by lower incentives, i.e., fewer credits for each patent examined). They show that promotions are associated with sharp increases in grant rates, controlling for examiner experience, which they interpret as less rigorous screening and lower quality patents. While this is a striking finding, their analysis does not determine whether it is driven by differences in extrinsic incentives, intrinsic motivation, or examiner opportunity costs. Our structural model, which embeds the strategic interaction between the applicant and examiner in a dynamic negotiation process, allows us to assess the separate impacts of these factors.

Perhaps the most closely related paper is Schankerman and Schuett (2022), who develop an integrated framework to study patent screening, encompassing the patent application decision, examination, post-grant licensing, and litigation in the courts. They calibrate the model on data for the U.S. and use it to evaluate various counterfactual patent and court reforms. Their model estimates the effectiveness of patent examination, which they treat as exogenous, but they do not model the examination process. Our paper complements their analysis as we develop the first equilibrium model of patent examination itself, which allows us to explain how reforms to the incentives and structure of screening affect patent quality. At the same time, however, our paper does not model the ex post screening of granted patents through licensing and litigation in the courts, which is a central focus of Schankerman and Schuett (2022).

Empirical Models of Bargaining

Finally, our paper is related to the literature on structural bargaining models. The final stage of our examination game is a model of negotiation between the patent applicant and examiner. The applicant requests a set of claims that define the scope of property rights, and the examiner either accepts the claims or requires some degree of narrowing of their content. This interaction proceeds in a multi-round setting. While this model is not exactly a “bargaining” framework, it is similar in structure. In most empirical studies of bargaining, for reasons of tractability, researchers have adopted the Nash model of bargaining (e.g., Grennan (2013) and Gowrisankaran, Nevo, and Town (2015)). In Nash bargaining, one does not specify the exact structure of negotiation, which is attractive in environments where the structure is unknown. We do not adopt Nash bargaining because it cannot accommodate negotiation breakdown, which is an essential feature of our environment, and the structure of negotiation in our environment is known and can be directly incorporated into our model.

3 The Patent Prosecution Process

We now briefly describe the patent prosecution process. We provide the minimal detail necessary to understand the general structure of our model. For more information, see Graham, Marco, and Miller (2018) and Merges and Duffy (2002).

A patent is a grant of the right to stop others from making, using, or selling an invention in a given jurisdiction for a limited period.[4] In exchange for these “monopoly” rights, the inventor is

[4] A patent does not preclude licensing to others, which occurs widely in the “market for technology.” Quite the contrary, patent rights underpin this market by helping to solve the informational difficulties in markets for knowledge. But a patent, particularly its scope, affects the outside options – and thus the division of the returns

meant to describe the full details of the invention in the patent document to promote information diffusion and facilitate later innovation building on the invention (so-called "enablement" requirement). The critical feature of the patent document is the list of claims. Claims delineate the "metes and bounds," or scope, of the property right (Merges and Duffy, 2002). Independent claims are the primary expression of the boundaries of patent rights and the main source of private value to the applicant. Dependent claims act more as clarifying devices, identifying successively narrower interpretations of the independent claim to which it refers. The examination process involves assessing the patentability of *each* claim, not the patent as a single entity.

In a departure from most existing literature, we treat a patent as a collection of claims that differ in scope rather than as a single entity. Claim heterogeneity has two dimensions: their similarity, or proximity, to previous patented claims and their private value. These two dimensions are a critical feature in the model. For many purposes, it is adequate to focus on the patent as the object, but we believe this heterogeneity is a first-order feature necessary to match the actual process of patent examination and to develop accurate statements about the potential effects of regime changes on the patent examination process. We embed heterogeneity in our model by endowing the inventor with multiple independent claims of varying distances to prior patents and private value.

3.1 Examiner Assignment and Search

After the application is submitted, the Patent Office assigns the application to an examiner in the relevant technology area, known as an art unit. Most art units randomly assign applications to examiners (Lemley and Sampat, 2012; Feng and Jaravel, 2019). The designated examiner searches the "prior art" – existing patents and non-patent literature in the public domain (e.g., scientific publications) – and decides whether to grant the claims in the patent. The examiner does this by checking whether the claims meet the legal standards of patentability. The three main grounds are novelty (35 U.S.C. §102), non-obviousness (35 U.S.C. §103), and indefiniteness (35 U.S.C. §112).[5] Novelty requires that the claim has not been in use for one year before filing.

from the innovation – in any licensing agreement. In doing so, the scope of patent rights affects the ex ante incentives to innovate.

[5] There is another ground for rejection – subject matter ineligibility (35 U.S.C. Section 101). To be eligible for a patent, the claimed invention must fall into one of four categories – process, machine, manufacture or composition of matter, as interpreted by the courts. Laws of nature, natural phenomena and abstract ideas have been deemed ineligible. Section 101 rejections account for 4.5% of all rejections, and are concentrated in software and business methods, though not exclusively. Because they only represent a relatively small fraction of rejections, we do not build them into our baseline economic model.

Non-obviousness requires that the claim makes an inventive step beyond the closest existing invention that would not be immediate to anyone skilled in the relevant area. Indefiniteness requires that the claim is precise and clear on the exact boundaries of claimed property rights. In this paper, we focus on novelty/non-obviousness.[6]

3.2 Negotiation

If the examiner grants a patent, the application process ends. If the examiner does not grant, they give a "non-final" rejection that informs the applicant which of the claims the examiner rejected and the grounds for the refusal. The applicant can amend and resubmit their claims. The examiner responds to the amended application similarly, except now a rejection is labeled "final." Despite the name, the applicant can respond *indefinitely* to "final" rejections by paying for Requests for Continued Examination (RCE). This negotiation lasts until the applicant abandons or the examiner grants the patent (a feature that differs from patent offices in other countries). We return to this point in our counterfactual analysis, where we analyze the consequences of limiting the number of allowable rounds of negotiation.

Examiners receive "credits" for their decisions during patent examinations. These credits contribute towards targets set by the Patent Office, which examiners are expected to meet. The allowed credits are adjusted by a seniority factor (senior examiners receive fewer effective credits for the same tasks) and according to the complexity of the examiner's technology area.[7] Also worth noting is that the credits decline over the various rounds of the examination process, with *fewer* credits for equivalent actions in RCEs relative to non-final and final rejection rounds.

3.3 Post-Grant Renewal

If the examiner grants a patent and the applicant pays the finalizing fees, the applicant receives legal protection over the invention. To maintain the enforceability of the patent, the applicant must pay renewal (or maintenance) fees before the fourth, eighth, and twelfth year after issuance. If not, the patent expires. The maximum statutory duration of patent rights is 20 years after the applicant *files* the patent.

[6] Using the Office Action Research Dataset described in Section 4, which identifies the reasons the examiner rejects claims in a patent at each round, we analyzed the overlap between novelty/non-obviousness (102/103) and indefiniteness (112) rejections. We find that 73% of office actions containing a 112 rejection also contain a 102/103 rejection. Thus, novelty/non-obviousness rejections cover most of the observed indefiniteness rejections, so omitting indefiniteness from the baseline model is a profitable abstraction.

[7] See Foit (2018) for further details and Appendix section E.2 for the way that we specify credits in our model.

4 Data and Descriptive Results

In this section, we describe our primary data sources, focusing on datasets not previously used in empirical studies of patents. We also present summary statistics and describe reduced-form evidence. Appendix B provides hyperlinks to all publicly available datasets and data sources we used in our empirical work.

Distance Metric

We construct a new measure of independent claim distance. To create this, we exploit the *U.S.PTO Patent Application Claims Full Text Dataset* and the *Granted Patent Claims Full Text Dataset*. The first dataset contains the full text for all U.S. patent application claims between 2001 and 2014 and an indicator for whether the claim is independent.[8] The Granted Patent Claims Full Text Dataset records the full text for all U.S. patent claims granted between 1976 and 2014.

We summarize our approach to creating a distance measure here; Appendix C provides a technical description of our methodology. To determine the distance between a patent application claim and the closest claim in existing patents, we compute the distance to every previously granted independent claim and then take the minimum. The approach calculates the distance by representing a patent claim's text as a numerical vector and calculating a metric on that vector space.[9] The standard method (*bag-of-words*) for representing the patent claim text as a numerical vector has two significant weaknesses: it ignores the *ordering* and *semantics* of words.[10] Instead, we use the *Paragraph Vector* approach of Le and Mikolov (2014). This approach uses an unsupervised algorithm to "learn" the meaning of words by studying the context in which they appear and forming a vector representation for each word, picking up the meaning of paragraphs as a by-product. The approach allows for synonyms, antonyms, and technical terminology with similar meanings, which are not accounted for by the bag-of-words method. As is common in the NLP literature, we compute distances between numerical vectors using the angular distance metric.[11]

[8] Our model and empirical work focus on independent claims because they are the primary determinant of the boundaries of the property right. In Appendix E.1.3, we show how the model could be extended to incorporate dependent claims. This extension would involve considerable complication.

[9] Kelly, Papanikolaou, Seru, and Taddy (2021) use similar methods to calculate patent similarity.

[10] For example, with bag-of-words, the sentences "The strong man lived near New York" and "A powerful male resided in the Big Apple" would both be closer to "The woman's favorite brand of phone is Apple" than to each other.

[11] We conduct two falsification tests on our distance measure. First, we put independent claims into twenty, five-percentile bins of the distance measure and then calculate the proportion of claims rejected on novelty/obviousness

Rounds Data

Since we estimate a model of the patent prosecution process over multiple rounds, comprehensive and reliable *round-level* data on the patent process are essential. We use the *Transactions History* data in the *Patent Examination (PatEx) Research Dataset* to create a dataset on the round-by-round evolution of utility patent applications between 2007 and 2014. The transactions dataset includes 275.6 million observations covering 9.2 million unique applications. For every patent application, these data record the type and date of every recorded written communication between examiner and applicant during the examination process. We extract the events corresponding to rejections, RCEs, abandonments, and grants to create a dataset on application outcomes by round.

Sources Matched to Round Data

We match the round-level data to three other datasets on patent applications. The first is the *Application Data* in the *PatEx* Dataset, which contains features of the patent application, such as the identities of the applicant and examiner, the patent art unit, and a binary indicator of the size of the applying firm (below or above 500 employees). We use these data to obtain the distribution of applications across art units and technology centers, the length of the examination, and the match between applications and examiners.

Second, we match our data to renewal decisions using the *U.S. PTO Maintenance Fee Events Dataset.* These data record the fee status of every granted patent since 1981, from which we calculate the renewal decisions. We estimate our model using the proportion of patents renewed at each required stage.

Third, since our model focuses on novelty/obviousness rejections, we require data on the *types* of rejections of each claim at each stage of the process. We obtain this from the *U.S.PTO Office Action Research Dataset for Patents* (Lu, Myers, and Beliveau, 2017). These data report all the grounds for rejection for rejected claims. From this, we calculate the overlap between the different grounds for refusal, which motivates our focus on novelty/obviousness rejections in the baseline model (see Subsection 3.1).

Legal Fees

grounds in each bin. We would expect that examiners are more likely to reject claims with a small distance to existing claims based on novelty/obviousness criteria. Thus the proportion of first-round rejections should be a declining function of the distance metric and the results confirm this prediction. Second, we conduct a similar test on the average number of examination rounds for each granted *patent*, by five-percentile bins of average distance of independent claims. Patents with higher average distance should be granted faster, and this is what we find.

For attorney fees, we use data from the *2017 American Intellectual Property Law Association (AIPLA) Report of the Economic Survey*. Based on the responses of about 450 intellectual property legal firms, the survey reports means and percentiles of the distribution of hourly fees for different tasks, such as preparing and filing an application, issuing, paying renewal fees, and amending applications. The survey separately reports this information by the complexity of the application ("simple" applications with fewer than ten claims and "complex" applications) in the biotechnology/chemical, electrical/computer, and mechanical fields. We use these moments to estimate the distributions of application and fighting costs for each patent application, adjusted for inflation.

Seniority and Technology Complexity Credit Adjustments

We obtained technology credit adjustments from the Patent Office at the disaggregated U.S. Patent Classification level and then aggregated them to the technology center level. Finally, we obtain data on examiner seniority from Frakes and Wasserman (2017), who provide a panel of General Schedule (GS) grades for examiners, including each examiner's promotion dates. Using this, we can work out the seniority of the examiner for each application.

4.1 Summary Statistics

Table 1 provides summary statistics on regular resolved utility patent applications filed between 2001 and 2017. Several features are worth noting. First, 70% of applications resulted in the issuance of a patent. However, this is a misleading measure of the fraction of *content* granted because, as we will see, most applications are heavily narrowed during the examination process. Second, the prosecution time varies across applications – the mean duration is 2.96 years, and the mean number of rounds is 2.40 (66% of applications concluded with two rounds and 83% within three). Third, the mean, median, and modal number of independent claims is three. Fourth, 24% of applications were by firms with fewer than 500 employees (a so-called "small entity"). Lastly, 46% of granted patents were renewed to the statutory limit, and only 13% were not renewed at the first renewal date.

Table 1. Summary Statistics

Variable	Observations	Mean	Median	Std. Dev.
Issued	4,846,053	0.70	1.00	0.46
Duration of Prosecution (years)	4,846,053	2.96	2.67	1.57
Number of Rounds	4,608,833	2.40	2.00	1.45
Independent Claims	3,838,553	2.99	3.00	2.94
Small Entity	4,781,012	0.24	0.00	0.43
Not Renewed at 4	410,667	0.13	0.00	0.33
Renewed at 4, not at 8	410,667	0.19	0.00	0.39
Renewed at 8, not at 12	410,667	0.23	0.00	0.42
Renewed at 12	410,667	0.46	0.00	0.50

Notes: Sample sizes are lower for rounds, claims, and examiner variables since the datasets containing these variables cover a subset of the years 2001-2017. On renewal variables, we restrict attention to patents granted before 2006 to ensure that we have full renewal data on all granted patents. Categorical variables may not sum to one due to rounding.

4.2 The Effects of Examiner Seniority and Technology

Existing studies show that patent grant rates vary widely across technology centers and examiner seniority, with more senior examiners granting more frequently (Frakes and Wasserman, 2017; Gaule, 2018; Sampat and Williams, 2019). In Appendix D, we confirm these findings about grant rates using our data, and we also show that the likelihood of multi-round negotiation (lasting beyond one round) is much lower for senior examiners and varies substantially across technology centers. In addition, small entities are less likely to negotiate. We also analyze the variation in these outcomes for *each* examiner, decomposing variation in examiner-specific outcomes (such as their grant rate) within and between technology center-seniority pairs.[12] This decomposition shows that 80% of the variation in examiner grant rates and 81% of the variation in each examiner's average number of rounds is within-group variation.

Our model allows for several factors that can explain the substantial variation in examiner statistics even within seniority-technology-center dyads: we allow for a different distribution of intrinsic motivation for junior and senior examiners, we incorporate differences in the examiner credit structure across seniorities and technology centers, and we allow for heterogeneous legal

[12] Table D.2 provides more detail, along with the proportion of within-group variation for other dependent variables, such as mean examination length, mean number of rounds, etc.

(fighting) costs for applicants across technology centers. Our parameter estimates will enable us to disentangle the effects of these factors in explaining the variation in outcomes.

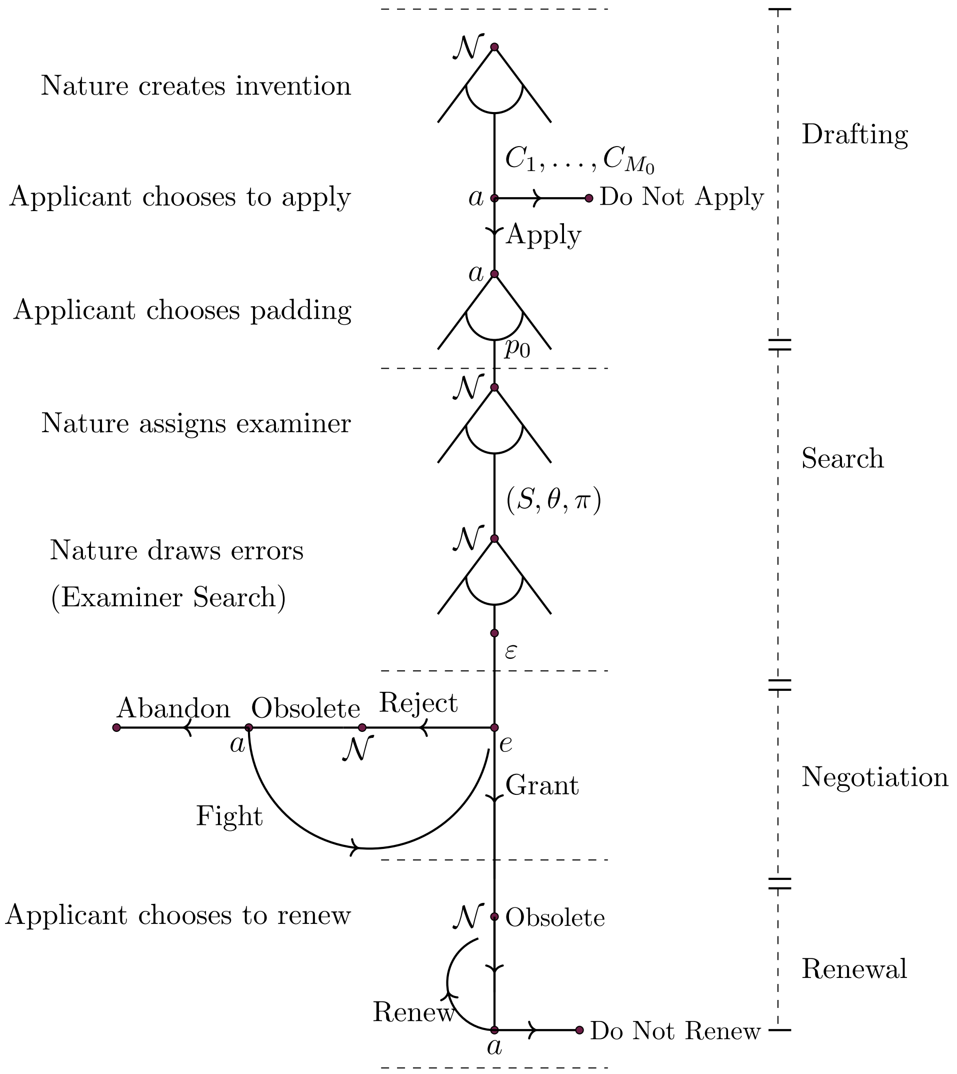
5 Model of the Patent Screening Process

We model the patent screening process as a dynamic game in technology center T, between an inventor, a, and a realization of the examiner, e. The game has four potential stages: (1) Application Decision and Patent Drafting, (2) Examiner Search, (3) Negotiation, and (4) Renewal. Figure 1 depicts the extensive form of the game.

In the baseline model, we analyze patent screening *conditional* on the invention being developed. For the validity of the structural model (and the counterfactual analysis), we do not need to model the potential inventor's decision whether to invest to develop their idea into an invention. However, to quantify the *net social costs* associated with these errors, we need to model the decision to develop (as well as how the patentee licenses their invention), which we do in Section 9.

Regarding the examiner, we present a model of how they act on *one specific application.* Therefore, we focus on intra-application incentives and costs for the examiner, rather than inter-application incentives induced by factors such as meeting their quarterly credit targets. A model in which examiners make decisions over time with consideration of the complete set of examinations in their docket would introduce significant complications and is not necessary to meet the aims of our model.

Figure 1. Extensive Form of the Model



5.1 Application Decision and Patent Drafting

5.1.1 Inventor Type

An inventor is endowed with a developed invention they are considering patenting. The patent application for the invention consists of M_0 initial independent claims $(C_1, ..., C_{M_0})$.[13] We characterize an independent claim C_j by the pair (D_j^*, v_j^*) where $D_j^* \sim G_D(\cdot)$ is the distance of the true version of claim j to the nearest claim in any *existing* invention and $v_j^* \sim G_v(\cdot)$ denotes the

[13] As we want to focus on the economic incentives for the applicant, we do not consider any agency issues between the inventor and the patent attorney who actually drafts the application.

initial flow net returns generated by the true version of claim j once it is commercialized.[14] We define the returns v_j^* as relative to the inventor's outside option, e.g., protecting the invention by trade secrecy.[15]

5.1.2 Application Decision

First, the inventor decides whether to apply. If they do not, the game ends, and their payoff is zero. If they do, they become an *applicant*, and the game continues. The inventor, a risk-neutral expected utility maximizer, chooses to apply if the expected utility of the game that follows applying is positive (because flow returns are defined relative to the next best alternative).

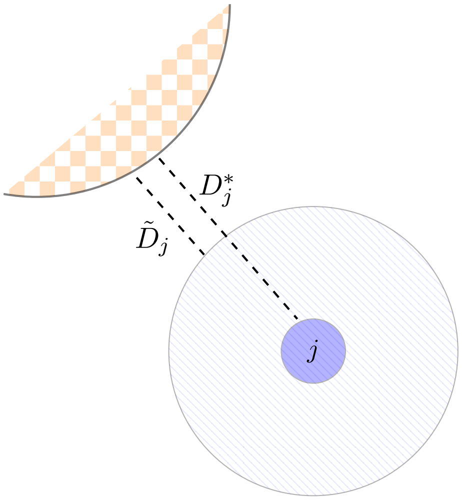
5.1.3 Padding

After deciding to apply, the applicant chooses the amount to exaggerate the claims on their patent application. We refer to this as the initial choice of *padding*, denoted p. Padding obfuscates the true "metes and bounds" of the invention, thereby concealing the inventive step and expanding the property rights claimed in the application. Padding allows the patent owner to extract potentially more revenue, by working it themselves or licensing it. However, greater padding also entails some obfuscation in defining the relationship between the actual invention and the boundaries of the patent rights claimed and necessarily moves the application closer to the prior art. Figure 2 illustrates the concepts of independent claims and padding.

[14] We assume that distances and values are uncorrelated. Based on the theoretical literature on differentiated products, the relationship is ambiguous. Other things equal, being further from rivals (in product space, which we assume is correlated with claim distances) softens price competition and thus increases private value – implying a positive correlation between distance and value. However, the distribution of demand will typically vary with location, with firms endogenously locating (patenting) in areas of high demand. This implies a negative correlation between distance to rivals and values.

[15] Table 2 provides our choices for parameterized distributions of distances $G_D(\cdot)$ and values $G_v(\cdot)$ (along with all other distributions in the model).

Figure 2. Distances and Padding



Notes: This figure provides a visual representation of independent claims and padding. The Figure is situated in the intellectual property "space". The orange semicircle in the top left corner represents the closest existing invention to the independent claim j, which is the small full blue circle in the bottom right corner. The applicant pads the true independent claim to create the larger cross-hatched circle. The distance between the true independent claim and the nearest existing invention is D_j^*, whereas the distance between the padded claim and nearest point is $\tilde{D}_j$.

There is a tradeoff for the applicant in the choice of padding. The advantage is that it increases the initial returns of claim j for the applicant from v_j^* to $\tilde{v}_j^0 = \mathcal{V}(v_j^*, p)$, where the padded value function $\mathcal{V}(\cdot, \cdot)$ is increasing in both arguments. On the other hand, padding increases the likelihood of examiner rejections during the examination process on the grounds of non-obviousness (closeness to existing patents) and indefiniteness. Padding shrinks independent claim distances from D_j^* to $\tilde{D}_j^0 = \mathcal{D}(D_j^*, p)$, where the padded distance function $\mathcal{D}(\cdot, \cdot)$ is increasing in D_j^* and decreasing in p. For simplicity, we assume that value (distance) is proportional (inversely proportional) to the degree of padding: $\tilde{v}_j^0 = p \cdot v_j^*$ and $\tilde{D}_j^0 = D_j^*/p$.

Finally, there is a direct cost of padding in the form of legal costs, which we assume is proportional to padding because heavily padded applications require more time to craft.[16] In particular, we

[16] The applicant may choose to *understate* the true scope of the invention ($p < 1$) and thus earn lower returns, as it reduces the likelihood of rejection by the examiner (especially if there is a restriction of the number of rounds allowed). We find some evidence of such under-padding in the empirical results.

specify legal costs as $F_{\text{app}} = f_{\text{app}} \cdot (1 + |p - 1|)$, where f_{app} is the attorney fees associated with patent drafting (which is log-normally distributed across applicants). The motivation for this specification is that it takes additional time for the attorney either to under-pad ($p < 1$) or over-pad ($p > 1$); writing down the truth ($p = 1$) is quickest. We assume symmetry for simplicity.

5.1.4 Applicant Expected Utility

The applicant decides the initial level of padding without knowing the identity of the examiner the Patent Office will assign. This feature is relevant because examiners differ in types (seniority, time cost, and intrinsic motivation) and, thus, in their strategies. As a result, applicants make initial padding decisions in light of the distribution of examiner types. The applicant chooses initial padding to maximize their expected utility less application legal costs, where the expectation is taken first over the roster of potential examiners $e = 1, \dots \underline{E}$ (where the random assignment of applications implies an equally likely chance of each examiner in the relevant technology center), over the error of the examiner $\varepsilon \sim G_{e,\varepsilon}(\cdot)$, and potential obsolescence of their invention $\boldsymbol{\omega}$ (all described later).

Formally, the applicant's optimal padding choice p_0 maximizes the ex ante value of patent rights $\Gamma(p)$, defined[17]

$$\Gamma(p) = \mathbb{E}_{e,\varepsilon,\boldsymbol{\omega}}\Big[U_a^0(e, \varepsilon, \boldsymbol{\omega}, p)\Big] - F_{\text{app}}(p),$$

where

$$\mathbb{E}_{e,\varepsilon,\boldsymbol{\omega}}\Big[U_a^0(e, \varepsilon, \boldsymbol{\omega}, p)\Big] = \frac{1}{\underline{E}} \sum_{e=1}^{\underline{E}} \int \mathbb{E}_{\boldsymbol{\omega}} U_a^0(e, \varepsilon, \boldsymbol{\omega}, p)\, dG_{e,\varepsilon}(\varepsilon),$$

and we define the applicant expected utility (over the full vector of obsolescence) for a given examiner e and error ε, denoted as $\mathbb{E}_{\boldsymbol{\omega}} U_a^0(e, \varepsilon, \boldsymbol{\omega}, p)$, later in Equation (4).[18] The applicant applies if

$$\Gamma^* \equiv \Gamma(p_0) \geq 0. \tag{1}$$

[17] Throughout, we use the notation $\mathbb{E}_{\boldsymbol{\omega}}$ to denote expectations taken over the vector of obsolescence shocks that are not yet realized. Before applying, this is the full vector of 20 possible shocks that could occur, one each year after application. As the process continues, obsolescence shocks occur, and fewer shocks are left to be realized. With a slight abuse of notation, whenever we use $\mathbb{E}_{\boldsymbol{\omega}}$ with an emboldened $\boldsymbol{\omega}$, it refers to the sub-vector of $\boldsymbol{\omega}$ that have not yet occurred.

[18] We simplify notation by using e to denote both the random variable reflecting the (unknown) examiner prior to application its realization after applying. The same holds for examiner errors.

5.2 Examiner Search

5.2.1 Examiner Assignment

The patent office assigns the application randomly to an examiner within the relevant art unit of the technology center. We characterize an examiner by the tuple (S, θ, π). The first term S represents examiner seniority. The type $\theta \sim G_{S,\theta}(\cdot)$ corresponds to the level of intrinsic motivation. More intrinsically motivated workers place a higher disutility on awarding patent rights that differ from their best estimate of the true scope embodied in the invention (see Section 5.4.3 for formalization of how this enters the examiner's payoff). We let the distribution of θ depend on seniority S. Finally, $\pi \sim G_{\pi}(\cdot)$ corresponds to the examiner's cost of delay (i.e., the extra effort cost for going another round plus any pressure costs associated with timely docket management). The effort cost component will reflect the examiner's productivity.

5.2.2 Examiner Grounds for Rejection

Once assigned, the examiner learns the applicant's identity and thus their fighting costs. The examiner also knows the padded value of the application to the applicant. The examiner reads the application and independently searches the existing prior art to assess the grounds for rejection throughout the negotiation process. We focus on the *obviousness/novelty* ground for rejection (35 U.S.C. §102/103).[19] After searching the prior art, the examiner assesses the obviousness/novelty of *each claim* j, denoted by $\hat{D}_j$ and equal to

$$\hat{D}_j = \mathcal{D}(D_j^*, p) \cdot \varepsilon,$$

where ε denotes the drawn examiner error in assessing obviousness/novelty, which is assumed to be independent of the true distance D_j^*. The distribution of search errors depends on the seniority of the examiner and may also depend on the technology center since the number and complexity of patents and other prior art vary across technology fields.

The distribution of search errors also depends on the intrinsic motivation of the examiner. We specify that the *mean* of the search error satisfies two criteria. The first is that the mean of the error tends to one (the unbiased case) as $\theta \to \infty$. The second is that for all $\theta < \infty$, the mean of the search error distribution is greater than one. We specify the second feature because examiners who are not perfectly intrinsically motivated do not scour the literature so thoroughly, thereby missing relevant prior art. When they miss relevant prior art, they perceive distances to

[19] In Appendix E.1.2 we describe an extension explaining how one could incorporate indefiniteness. See footnote 6 for an empirical justification on why we abstract from indefiniteness (35 U.S.C. §112) and subject matter ineligibility (35 U.S.C. §101) in our main specification.

be larger than they are and hence have errors greater than one. However, these requirements do not force one-sided examiner error since some draws may still be below one, even if the mean is above one. Our functional form choice satisfying this assumption is $\mu_\varepsilon = 1 + \frac{1}{\theta}$.

We say the examiner has *grounds* for an obviousness rejection if $\hat{D}_j$ is less than an obviousness threshold τ. However, having grounds for rejection will not necessarily mean the examiner will reject the claim. The examiner's decision will be the one that maximizes their utility, taking into account their explicit incentives (credits) and intrinsic motivation. This is a crucial point as it implies that examiners' decisions in the data may not align with decisions made solely on legal grounds.

Finally, examiner errors are specified to be constant throughout the negotiation stage. In this sense, there is no updating of examiner error.[20] However, the grounds for rejection *will* be recalculated at every negotiation round as the applicant narrows the extent of padding in response to a rejection by the examiner.

5.3 Information Structure

The information structure for the applicant and examiner is as follows. The inventor knows the set of claims covered by the invention (given by nature), their true distance to all prior art, and the private value of each claim. Before deciding whether to apply for a patent, the inventor does not know which examiner will be assigned to the application. After assignment, the applicant knows the characteristics of the examiner, including the level of intrinsic motivation, seniority, and structure of patent office incentives the examiner faces. The applicant also knows the structure of the process and the fees imposed by the patent office at each stage.

The assigned examiner does not observe the true claim distances or the applicant's extent of padding, only the padded distances, contaminated by examiner error, for each claim in the application. The examiner does not know the error she makes in determining the claim distances during the search of prior art. The examiner observes the fighting costs and padded private value of the applicant's claims.[21] Since the examiner reports their assessment of the padded distance to the applicant, the applicant knows the examiner's error.

[20] See Appendix Section E.1.1 for a short discussion of how learning could be incorporated in an extended model.

[21] We could assume that the examiner does not perfectly observe the private value, but instead obtains an unbiased signal of the value. This feature would not deliver any additional insights and would increase computational burden.

5.4 Negotiation

The Negotiation Stage is a finitely repeated version of the stage game shown in the "Negotiation" section of Figure 1. At round r, if required to act, first, the examiner chooses whether to grant or abandon and, if rejected, the applicant chooses whether to abandon or fight. In between the examiner's and applicant's decision, the applicant's invention can become obsolete, in which case the applicant abandons it immediately. The applicant and examiner discount each stage at rate β.

Let $\mathbf{x}_a$ and $\mathbf{x}_e$ be the strategies of the applicant (including renewal decisions) and examiner, respectively, if the invention is not obsolete.[22] We detail the actions and payoffs obtained at the two decision nodes, starting at the point at which the examiner has just rejected in round r so that $x_e^r = \text{REJ}$.

5.4.1 Obsolescence and Credits

First, pre-grant obsolescence, denoted by ω_r, is realized. If $\omega_r = 1$, the applicant's invention becomes obsolete. In this case, all returns shrink to zero permanently, and trivially, the applicant abandons and obtains a period payoff of zero.[23] In this case, the examiner obtains a period payoff of credits $g_{ABN}^r(S, T)$. If the invention does not become obsolete, then $\omega_r = 0$, and the applicant makes a non-trivial decision. Formally, obsolescence is a Markov process, where, for all r, if $\omega_r = 1$, then $\omega_{r+1} = 1$ (an absorbing state). Otherwise, if $\omega_r = 0$, ω_{r+1} is a Bernoulli random variable with parameter $P_{\omega,\text{pre}}$ if we are still in the application process, and parameter $P_{\omega,\text{post}}$ if a patent has been granted and we are in the renewals process.

We provide the full schedule of examiner credits in Appendix E. The most important feature to note is that credits weakly decline as the applicant enters subsequent requests for continued examination, which make early granting more attractive to the examiner.

5.4.2 Applicant Decision

Upon receiving a rejection, if the invention has not become obsolete, the applicant has two choices. They can abandon ($x_a^r = \text{ABN}$), in which case the applicant's and examiner's payoffs

[22] Of course, the vectors include a rejection/acceptance decision and abandonment/fight decision for *every* round. To check whether a strategy is optimal, we must specify what each player would do in every round, even if the prior parts of the strategy dictate that this round will not be reached on the equilibrium path.

[23] The applicant obtains a period payoff of zero because the Patent Office reveals all applications (after 18 months), so their potential for appropriation of innovation returns (e.g., by trade secrecy as an alternative) has essentially vanished.

are as described in the event of obsolescence. Instead of abandoning, the applicant can continue the application ($x_a^r = \text{FIGHT}$). Continuing involves narrowing rejected claims, which we model as a reduction in padding p by proportion η.[24] Hence for all *rejected* claims j, the padding becomes $p_{j,r+1} = \eta p_{j,r}$. The padding level remains the same for all *accepted* claims.

Continuing involves a fighting cost to the applicant. The applicant must pay the attorney the fee for amending the application, F_{amend}. In the case of a Request for Continued Examination, the applicant must pay the associated patent office fee, F_{round}^r. Continuation involves delay costs, denoted by π. After narrowing occurs, the applicant pays fighting costs, and we move to round $r+1$.

Formally, let the value function for the applicant *upon being rejected in round* r be $U_a^r(\omega_r, \mathbf{x}_e)$. Clearly, the value function for the applicant is a function of the future actions of the examiner. Further, because ω is a Markov process, the value function for the applicant only depends on the realization of ω in period r. The term $U_a^r(\omega_r, \mathbf{x}_e)$ is defined as follows. If the invention becomes obsolete, so that $\omega_r = 1$, we have (for all $\mathbf{x}_e$)

$$U_a^r(1, \mathbf{x}_e) = 0. \tag{2}$$

Otherwise,

$$\begin{aligned} U_a^r(0, \mathbf{x}_e) = \max\Bigg\{0, -F_{\text{amend}} &- F_{\text{round}}^{r+1} + \beta\Big(1(x_e^{r+1} = \text{GR})[V^{r+1} - \phi] \\ &+ 1(x_e^{r+1} = \text{REJ})\mathbb{E}_{\omega_{r+1}} U_a^{r+1}(\omega_{r+1}, \mathbf{x}_e)\Big)\Bigg\}, \end{aligned} \tag{3}$$

where $1(A)$ is the indicator function, equal to one if statement A is true and zero otherwise, V^{r+1} defines the ex post net expected benefits from patent rights if granted in round $r+1$, as given in Equation (10) in Section 5.5, and ϕ is the finalizing fee. Equation (3) says that the value for the applicant in round r, provided they are not obsolete, is either zero if it is optimal for them to abandon or the sum of fighting costs, plus either the payoff of being granted in the next round (if the examiner will grant them) or the expected value from round $r+1$ if the examiner will reject them in round $r+1$ (both discounted by β).

[24] We could extend the model to allow the applicant to choose whether to narrow by proportion η with some probability or respond by arguing that the examiner is in error and not narrow at all. However, our data on patent word counts imply that this extension is empirically unimportant. To see this, we look at word counts on patents granted with one rejection after publication and calculate the proportion of cases whether the applicant resubmits an application with the same word count. This happens only 7% of the time, so we view the choice to ignore the possibility of no narrowing as a profitable abstraction in the baseline.

If $x_e^{r+1} = \text{GR}$, and $\omega_r = 0$, the applicant abandons in round r if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta[V^{r+1} - \phi]$$

and if $x_e^{r+1} = \text{REJ}$, the applicant abandons in round r if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta\mathbb{E}_{\omega_{r+1}} U_a^{r+1}(\omega_{r+1}, \mathbf{x}_e).$$

At this point, we can define the expected utility for the applicant before applying, for a given choice of padding, as

$$\mathbb{E}_{\boldsymbol{\omega}} U_a^0(e, \varepsilon, \boldsymbol{\omega}, p) = 1(x_e^{1,*} = \text{GR})[V^1 - \phi] + 1(x_e^{1,*} = \text{REJ})\mathbb{E}_{\omega_1} U_a^1(\omega_1, \mathbf{x}_e^*), \tag{4}$$

where all four terms on the right-hand side are (implicitly) functions of the level of padding.

5.4.3 Examiner Grant/Rejection

If the applicant fights ($x_a^r = \text{FIGHT}$), we move to a new round $r+1$, and the examiner obtains updated assessments $\hat{D}_j^{r+1} = \mathcal{D}(D_j^*, p_0\eta^r)$ on previously rejected claims. Based on their updated assessment, the examiner recalculates the grounds for rejection and decides whether to grant the patent.

Granting

Granting a patent in round $r+1$ ($x_e^{r+1} = \text{GR}$) ends the negotiation game and moves the applicant into the renewal stage. Let $\mathcal{R}^{r+1} \in [0, 1]$ denote the proportion of claims the examiner thinks they should reject on obviousness/novelty grounds. Then the immediate payoff to the examiner from granting is

$$\mathcal{G}^{r+1} = g_{GR}^{r+1}(S, T) - \theta\mathcal{R}^{r+1}.$$

Here $g_{GR}^{r+1}(S, T)$ is the credit received by the examiner for granting at stage $r+1$. The term $\theta\mathcal{R}^r$ captures the intrinsic utility cost for the examiner. For intuition on this term, consider the extreme cases. When $\mathcal{R}^{r+1} = 0$, the examiner believes there are no independent claims on which they have grounds to reject and therefore feels no intrinsic disutility in granting the application. On the other hand, when $\mathcal{R}^{r+1} = 1$, the examiner believes that they should reject every independent claim, so the examiner is going against the organization's mission statement in granting a patent. The examiner's intrinsic penalty from premature granting is the product of the proportion of strategically incorrect claim acceptances and their intrinsic motivation parameter.

Rejecting

If the examiner chooses not to grant in round $r+1$ ($x_e^{r+1} = \text{REJ}$) they get credits $g_{REJ}^{r+1}(S, T)$, and the stage game continues. The examiner follows this choice by rejecting any claim on which there are grounds to reject. Hence, the examiner rejects any independent claim j if $\hat{D}_j^{r+1} < \tau$. After this, the application moves back into the hands of the applicant, at which point another obsolescence realization occurs, and then the applicant decides again whether to abandon or continue.

Formally, we define the value function for *the examiner after rejecting in round* r as $W_e^r(\omega_r, \mathbf{x}_a)$. If the invention becomes obsolete, $\omega_r = 1$, and

$$W_e^r(1, \mathbf{x}_a) = g_{ABN}^r. \tag{5}$$

for all $\mathbf{x}_a$. Otherwise, if $\omega_r = 0$, the value function for the examiner satisfies

$$W_e^r(\omega_r, \mathbf{x}_a) = \begin{cases} g_{ABN}^r & \text{if} \quad x_a^r = \text{ABN} \\ -\pi + \beta \max\left\{\mathcal{G}^{r+1}, g_{REJ}^{r+1} + \mathbb{E}_{\omega_{r+1}} W_e^{r+1}(\omega_{r+1}, \mathbf{x}_a)\right\} & \text{if} \quad x_a^r = \text{FIGHT} \end{cases} \tag{6}$$

In the bottom branch of Equation (6), where the applicant fights, the value to the examiner of rejecting in round r is the cost π plus either the (discounted) benefits of granting in round $r+1$ or the net benefits of rejecting in round $r+1$, whichever is larger.

Given the applicant's strategy $\mathbf{x}_a$ the examiner grants in round r if

$$\mathcal{G}^r > g_{REJ}^r + \mathbb{E}_{\omega_r} W_e^r(\omega_r, \mathbf{x}_a).$$

This says that the examiner grants if the period payoff from granting exceeds the credits from rejecting plus the expected continuation value from the point of having rejected in round r, with expectation taken over obsolescence outcomes.

5.5 Renewal

We enter the renewals stage if the examiner grants the patent and the applicant pays the finalizing fee. Our renewal model adapts Schankerman and Pakes (1986) to the United States context, adding a probability of post-grant obsolescence in addition to deterministic depreciation. Suppose the patent is granted in round r. The returns for each granted claim j start at $\tilde{v}_{j,r} = v_j^* \cdot p_r$ and depreciate at rate δ each period after grant. With probability $P_{\omega,\text{post}}$, the invention becomes obsolete, at which point the returns shrink to zero permanently. To keep the patent rights, the applicant must pay renewal fees F_4, F_8, and F_{12} at years four, eight, and twelve after grant. The patent life ends at $L = 20$ years after submission of the patent application, at which point the invention enters the public domain.

The renewal decisions by the applicant are those that maximize their expected utility from retaining patent rights. Formally, define the expected returns from years t_1 to t_2 as

$$\mathbb{E}_{\boldsymbol{\omega}} V_{t_1,t_2} = \sum_{t=t_1}^{t_2} [\beta(1-\delta)(1-P_{\omega,\text{post}})]^{t-t_1} \sum_j \tilde{v}_{j,r}$$

and let I_t be equal to one if the applicant will renew at year t (provided the patent is not obsolete) and zero otherwise. Then, the applicant will renew at year four if the net expected benefit after year four is positive:

$$V_4^{N,r} \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{4,7} - F_4 + I_8 \beta^4 V_8^{N,r} > 0, \tag{7}$$

where $V_8^{N,r}$ is the net returns from patent rights after year eight, which is defined analogously. The renewal decision at year eight is analogous, and the decision at year 12 is similar, except there is no future renewal decision post year 12.[25] Finally, we define the ex post net expected benefits from patent rights, when granted in round r, denoted as V^r (as in Equation (3), as

$$V^r = \mathbb{E}_{\boldsymbol{\omega}} V_{1,3} + I_4 \beta^4 V_4^{N,r}. \tag{10}$$

5.6 Equilibrium

For every given parameter vector and choice of padding, the negotiation game is a finite game of perfect information, and hence has a subgame-perfect equilibrium that can be found through backward induction. The equilibrium strategies are characterized by (for all r):[26]

1. $x_e^{r,*} = \text{GR}$ if and only if
$$\mathcal{G}^r > g^r_{REJ} + \mathbb{E}_{\boldsymbol{\omega}} W_e^r(\omega_r, \mathbf{x}_a^*).$$
2. If $x_e^{r+1,*} = \text{GR}$, $x_a^{r,*} = \text{ABN}$ if and only if
$$F_{\text{amend}} + F^{r+1}_{\text{round}} > \beta[V^{r+1} - \phi].$$

[25] To be precise, conditional on not becoming obsolete, the applicant renews at year eight if

$$V_8^{N,r} \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{8,11} - F_8 + I_{12} \beta^4 V_{12}^{N,r} > 0, \tag{8}$$

and, conditional on not becoming obsolete, the applicant renews at year 12 if

$$V_{12}^{N,r} \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{12,20-r} - F_{12} > 0. \tag{9}$$

[26] In practice, we limit the process to six rounds (around 95% of applications last at most three rounds of negotiation and the modal number is two) so the characterization holds for $r < 6$. In the sixth round, if rejected, we force the applicant to abandon. The examiner's continuation value is therefore only $g^6_{ABN}(S,T)$.

3. If $x_e^{r+1,*} = \text{REJ}$, $x_a^{r,*} = \text{ABN}$ if and only if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta \mathbb{E}_{\omega_{r+1}} U_a^{r+1}(\omega_{r+1}, \mathbf{x}_e^*).$$

4. The terms $U_a^r(1, \mathbf{x}_e^*)$, $U_a^r(0, \mathbf{x}_e^*)$, $W_e^r(1, \mathbf{x}_a^*)$, and $W_e^r(0, \mathbf{x}_a^*)$ satisfy Equations (2), (3), (5), and (6), respectively.
5. I_4, I_8, and I_{12} are equal to one if and only if inequalities (7), (8), and (9), respectively, are satisfied.

Before turning to the estimation of the model, we want to highlight the important advantages of modelling patents as comprised of multiple claims rather than as a single object. First, since the vast majority of applications contain multiple independent claims, a model with multiple claims allows a more realistic description of the patent prosecution system and thus a tighter link to the data on which the model parameters will be estimated. Second, as described in Section 5.4.2, endowing applicants with multiple claims allows for specific claims to be narrowed only up to the round at which they are granted, which would not be possible in a model with a single claim. Third, as described in Section 5.4.3, including multiple claims allows for a more flexible and less discrete definition of the examiner's intrinsic motivation disutility – in particular, the examiner's cost of granting knowingly invalid claims is not binary but instead depends on the proportion of granted claims that are invalid. Finally, including multiple claims allow us to estimate intensive margin of examiners. For example, in the case of false grants, we will calculate the equilibrium proportion of granted claims that are invalid, in both baseline and counterfactuals, which would not be possible in a single-claim model.

6 Estimation

Our primary estimation method is simulated method of moments (SMM), though we estimate some parameters outside the model. For reference, Table 2 summarizes all parameters and their associated distributional assumptions.

Table 2. Estimated and Assigned Parameters

Estimated Parameters

Variable	Notation	Distribution	Parameters
Examiner			
Intrinsic motivation	$\theta \sim G_{S,\theta}(\cdot)$	Log-normal	σ_θ, $\mu_{\theta,\text{junior}}$ or $\mu_{\theta,\text{senior}}$
Examiner Delay Cost	$\pi \sim G_\pi(\cdot)$	Log-normal	μ_π, σ_π
Error	$\varepsilon \sim G_{e,\varepsilon}(\cdot)$	Normal	σ_ε
Applicant			
Initial claim returns	$v_j^* \sim G_v(\cdot)$	Log-normal	μ_v, σ_v
Initial claim distances	$D_j^* \sim G_D(\cdot)$	Beta	α_D, β_D
Obsolescence	$\boldsymbol{\omega}$	Bernoulli	$P_{\omega,\text{pre}}$ or $P_{\omega,\text{post}}$
Application legal costs	f_{app}	Log-normal	$\mu_{f,\text{app}}, \sigma_{f,\text{app}}$
Issuance legal costs	f_{iss}	Log-normal	$\mu_{f,\text{iss}}, \sigma_{f,\text{iss}}$
Maintenance legal costs	f_{main}	Log-normal	$\mu_{f,\text{main}}, \sigma_{f,\text{main}}$
Amendment legal costs	f_{amend}	Log-normal	$\mu_{f,\text{amend}}, \sigma_{f,\text{amend}}$
Narrowing	η	-	-

Assigned Parameters

Variable	Notation	Values
Discount rate	β	0.95
Depreciation	δ	$\frac{0.14-P_{\omega,\text{post}}}{1-P_{\omega,\text{post}}}$
Threshold by technology center	τ	Range from 0.48 to 0.52
Credits	$g^r(S,T)$	-
Finalizing fee	ϕ	\$2,268
RCE fees	$F^3_{\text{round}} = F^5_{\text{round}}$	\$1,034
Renewal fees	F_4	\$1,685
	F_8	\$3,791
	F_{12}	\$7,792

Notes: μ_ε is not estimated, instead it is set at $1 + \frac{1}{\theta}$. Legal costs differ by complexity and technology area. Composite depreciation and obsolescence are constructed from estimates in Bessen (2008); see Equation (11) for explanation. Threshold vector ranges from 0.48 to 0.52 across seven technology centers. Credits are provided in Appendix Section E.2. All fees are provided in 2018 U.S.D. Fees are 50% for small entities, which we incorporate in the model through assigning small entity status to 24% of applications (matching the proportion in Table 1).

6.1 External Estimation

Discount Rates (β_a, β_e)

The data lack some detailed information for identifying all parameters. Specifically, discount rates are traditionally difficult to identify. Hence, we fix them to $\beta_a = \beta_e = 0.95$—as in most of the literature (Pakes, 1986). As detailed in Section 7.4, we experimented with β equal to 0.99, and the results are quantitatively unchanged.

Distance Threshold (τ)

We estimate the distance threshold externally using observations on claim distances and examiners' grant decisions. For every examiner e, we calculate the minimum value of the distances among claims they grant. This number corresponds to their "personal distance threshold," denoted as $\tau_e = \min_{j \in J_e} \tilde{D}_j$, where J_e is the set of claims granted across all applications by examiner e. Since examiners are not perfectly intrinsically motivated, some examiners' personal thresholds are below the true threshold τ in cases where they knowingly grant patents with relatively small distances. However, we assume that the most intrinsically motivated examiner will have a personal threshold τ_e equal to the true threshold τ. This assumption allows us to estimate the distance threshold as the maximum of the distribution of examiners' individual thresholds. The validity of this approach relies on $\max\limits_{e=1,\ldots,\underline{E}} \tau_e \to \tau$ as the number of examiners in the technology center $\underline{E} \to \infty$.[27] We calculate these thresholds separately for each technology center to create technology-center-specific thresholds. Notably, our estimates of the threshold in different technology centers are very similar, ranging from 0.48 to 0.52.

Applicant Fighting Costs $(f_\cdot)$

We have data on the quantiles of the distributions of *amendment*, *maintenance*, and *issuance* hourly fees charged by lawyers. We assume these three costs are log-normally distributed. Since these moments directly correspond to the elements of applicant fighting costs and do not identify any other parameters in the model, we estimate the mean and variances of the log of fighting costs using the optimal two-step generalized method of moments estimation procedure for each of the three negotiation-based fighting costs.[28] Also, we observe different distributions for different technology areas, so we estimate different negotiation fighting cost distributions for *simple* applications (less than ten claims) and complex applications in *chemical*, *electrical*, and *mechanical* fields.[29]

[27] In practice, we experiment with the first and fifth percentiles as robustness checks in Section 7.4. We also remove examiners who have conducted fewer than a threshold number of examinations. We experiment with values of 50 and 100 for this threshold and find minor differences in each case.

[28] All three cases are over-identified because there are two parameters of a log-normal distribution, yet we observe the mean, median, and four percentiles of legal fees.

[29] On application fighting costs, though we have similar moments on lawyers' application drafting fees, because application fighting cost is proportional to padding in the model, its distribution is contaminated by the endogenous choice of padding (which is a function of all model parameters). This feature means that we cannot estimate the distribution of application fighting costs outside the model: we must estimate these parameters as part of the

Depreciation of patent returns (δ)

We do not directly estimate the depreciation rate for returns. Instead, we adapt the estimates from Bessen (2008) to our context. Bessen (2008) estimates the combined effect of depreciation and the probability of obsolescence as 0.14. As a result,

$$0.14 = (1 - P_{\omega,\text{post}}) \cdot \delta + P_{\omega,\text{post}} \cdot 1 \tag{11}$$

Hence, for each parameter guess of $P_{\omega,\text{post}}$, we extract the implied pure depreciation value by inverting Equation (11) to obtain the pure depreciation. See Schankerman and Schuett (2022) Appendix G for further information.

6.2 Simulated Method of Moments

We estimate the remaining set of model parameters using SMM. The model does not admit an analytic solution for endogenous variables as a function of all the model primitives. Hence, the goal is to choose the parameters that best match the moments of the data with the corresponding moments computed from the model's numerical solution. We estimate the model using moments from the data described in Section 4, assuming the model's equilibrium generates the data. We estimate pooled parameters across technology centers, though an extension estimating the model for each technology center is possible.

We denote the full vector of parameters to estimate as $\boldsymbol{\psi} = (\boldsymbol{\psi}_e, \boldsymbol{\psi}_a)$, composed of examiner and applicant parameter vectors $\boldsymbol{\psi}_e$ and $\boldsymbol{\psi}_a$, respectively. The vector of applicant parameters is $\boldsymbol{\psi}_a = \left(\eta, P_{\omega,\text{pre}}, P_{\omega,\text{post}}, \alpha_D, \beta_D, \mu_v, \sigma_v, \boldsymbol{\mu}_{\boldsymbol{f}_{\text{app}}}, \boldsymbol{\sigma}_{\boldsymbol{f}_{\text{app}}}\right)$. We described the narrowing and obsolescence parameters in Section 5. For distances, we assume D_j^* is Beta distributed with parameters (α_D, β_D). The Beta distribution is a natural choice as it provides a flexible distribution on the interval $[0, 1]$, which coincides with the interval of our distance metric. Further, we use a multivariate normal distribution copula to correlate claim distances within an application.[30] Motivated by Schankerman and Pakes (1986), the log of initial *claim* flow returns is normally distributed with mean μ_v and variance σ_v^2. Finally, we assume that the log of application drafting legal fees per unit padding, f_{app}, are normally distributed with mean $\mu_{f_{app}}$ and variance $\sigma^2_{f_{app}}$, with

simulated method of moments procedure described in the next subsection.

[30] Specifically, in the simulation, for each application, we draw a vector of size M_0 from a standard multivariate normal with correlation coefficient ρ. We apply the quantile function of the normal to the draws to create correlated uniform random variables. Then for the estimation guess $(\tilde{\alpha}_D, \tilde{\beta}_D)$, we apply the inverse CDF of a Beta distribution with these parameters to the uniform draws to generate correlated beta distributed initial distances. For ρ, we use the empirical correlation of granted distances. Simulations confirm that the correlation of the multivariate copula is very close to the correlation of the distances. See Nelsen (2007) for details.

different parameters for simple and complex applications in *chemical*, *electrical*, and *mechanical* fields.

The vector of examiner parameters is $\boldsymbol{\psi}_e = (\mu_{\theta,\text{junior}}, \mu_{\theta,\text{senior}}, \sigma_\theta, \mu_\pi, \sigma_\pi, \sigma_\varepsilon)$. The first three parameters $(\mu_{\theta,\text{junior}}, \mu_{\theta,\text{senior}}, \sigma_\theta)$ correspond to log-normal parameters for the distribution of examiner intrinsic motivation. We estimate different μ parameters for "junior" (pre-GS-14 grade) and "senior" examiners. Though we constrain the σ parameter to be the same for juniors and seniors, given the log-normal specification, this does not force the variances (or even the variance relative to the mean) to be the same for juniors and seniors. The log of examiner delay costs, π, are normally distributed with mean μ_π and variance σ_π^2. Finally, examiner errors are normally distributed, with variance σ_ε^2.[31]

We estimate $\boldsymbol{\psi}$ using a minimum-distance estimator that matches moments of the data with the corresponding moments implied by the model. More specifically, for any value of $\boldsymbol{\psi}$, we solve the model for several simulated draws from the distributions of exogenous variables. Then, we calculate moments of the endogenous variables across the simulated observations. The following section describes the moments we use and how they contribute to parameter identification.

The minimum-distance estimator minimizes the SMM objective function:

$$\hat{\boldsymbol{\psi}} = \arg\min_{\psi} \left(\boldsymbol{m}(\psi) - \boldsymbol{m}_{\mathcal{S}}\right)' \Omega \left(\boldsymbol{m}(\psi) - \boldsymbol{m}_{\mathcal{S}}\right),$$

where $\boldsymbol{m}(\psi)$ is the vector of simulated moments computed from the model when the parameter vector is ψ, $\boldsymbol{m}_{\mathcal{S}}$ is the vector of corresponding sample moments, and Ω is a symmetric, positive-definite weighting matrix.[32] We describe computational details in Appendix F.

6.3 Choice of Moments

We now briefly describe our choice of moments in $\boldsymbol{m}_{\mathcal{S}}$ for the SMM estimation. In Appendix G, we provide some intuition about how these moments aid in identifying the parameters we estimate.

The number of moments we can calculate on endogenous variables in the model far exceeds the number of model parameters. To select a subset of moments for our estimation procedure,

[31] As discussed in the model, the mean of each examiner's error distribution is not parametrized, instead linked to intrinsic motivation through the equation $\mu_\varepsilon = 1 + \dfrac{1}{\theta}$.

[32] For the weighting matrix we use a diagonal matrix that scales moments to a uniform scale. We cannot use the optimal two-step weight matrix because we do not have application-specific data on fighting costs that can allow us to compute the correlation between these moments and others.

we followed a rigorous, data-driven methodology, relying on three objects: (1) the sensitivity matrix of Andrews, Gentzkow, and Shapiro (2017) along with (2) Moments plots and (3) SMM plots, both described in Jalali, Rahmandad, and Ghoddusi (2015). We provide details on the complete set of moments we considered and our pruning procedure in Appendix G. Through this procedure, we pruned the set of moments down to 40 which assist in estimating the parameters.

The selected moments corresponding to outcomes for examiners are the proportion of applications granted by round and seniority, the standard deviation of examiner rejection rates by seniority, and the proportion of patents granted containing an invalid claim (again, by seniority and round). The selected moments corresponding to outcomes for applicants are the proportion of abandonments by round and examiner seniority, patent renewal rates, means and standard deviations of granted claim distances by round granted, and means and medians of legal application fees by technology class. Appendix G provides the exact set of moments in full detail.

7 Estimates, Robustness, and Fit

In this section, we present and interpret our parameter estimates and discuss model fit and robustness. In what follows, we bootstrap standard errors. Standard errors are negligible for all parameters, which is unsurprising since we calculate data moments using millions of observations.

7.1 Applicant Parameters

Table 3 presents the estimates for parameters relating to the applicant. First, we estimate the proportion of narrowing per round as $1 - \eta = 0.25$. This estimate indicates that screening substantially narrows over-claiming by the applicant. Second, we estimate two probabilities of obsolescence: a pre-grant probability during the application process and post-grant obsolescence during the patent's life. The estimated pre-grant obsolescence probability is 14% for each negotiation round. The post-grant rate is 4% per year, similar to other estimates in the literature. [33] The probability of obsolescence is higher during the application process for two reasons. First, applicants are more likely to discover their invention to be obsolete earlier in its life cycle (e.g., discovering that commercialization costs make the project unviable). Second, the prosecution stage contains applications that are eventually granted and those who abandon, and many of those who abandon do so precisely because they become obsolete.

[33] Using German data from 1953-1988, Lanjouw (1998) estimates a range of 7-12% for the post-grant obsolescence probability. Pakes (1986) calculates values of 6%, 4% and 1% for the likelihood of obsolescence in the first, second and third year after grant, respectively.

Table 3. Applicant Parameter Estimates

Parameter	**Symbol**	**Estimate**	**S.E.**
Per-round narrowing	η	0.75	0.000
Pre-grant obsolescence	$P_{\omega,\text{pre}}$	0.14	0.001
Post-grant obsolescence	$P_{\omega,\text{post}}$	0.04	0.000
Initial returns log-mean	μ_v	10.55	0.077
Initial returns log-sigma	σ_v	1.32	0.022
Initial distance alpha	α_D	4.57	0.003
Initial distance beta	β_D	7.74	0.004
Simple application fighting cost log-mean	$\mu_{f,\text{simple}}$	8.53	0.011
Simple application fighting cost log-sigma	$\sigma_{f,\text{simple}}$	0.87	0.054

Notes: This table provides the applicant's model parameters. Per-round narrowing is $1 - \eta = 0.25$. Standard errors are bootstrapped. Table A.1 provides fighting cost parameters by technology area.

Third, the distribution of initial returns from an unpadded independent claim is highly skewed. Though the mean is \$91,046, the median is \$38,069, and the modal value of initial returns for an unpadded independent claim is \$6,656. To understand the distribution of unpadded initial returns on the *application*, we take the distribution of the number of independent claims and use it to construct sums of draws from the distribution of claim returns. For example, the first patent application in our dataset has two independent claims. Hence, we draw two values from the distribution of claim initial unpadded returns and add them to get the total initial unpadded returns on that application. The median initial unpadded returns from a patent application are \$129,659.

It is difficult to compare our estimates of initial returns to existing estimates in the literature on total patent returns since we estimate the distribution of initial returns for (a) all *applications* (not just granted ones) and (b) *unpadded* claims. Nonetheless, it is worth noting that Bessen (2008) estimates the mean net present value of patents (adjusted to 2018 U.S.D) for all U.S. patentees as \$78,168 and \$113,067 for just U.S. public firms in manufacturing.

Next, we discuss the implied distribution of initial unpadded distances and fighting costs. The mean distance is 0.37, and the distribution is approximately symmetric. These estimates indicate that about 83% of application claims have distances below the threshold. Despite this, many applications are eventually granted because of extensive narrowing and examiners granting invalid claims. Fighting costs for simple applications are lower than all other categories. At the mean value of padding and application cost, we estimate simple application costs to be \$7,920 and \$12,333 for complex electrical applications.

Padding (overclaiming property rights)

We compute statistics on the model's endogenous variables by simulating the model at our estimates. Relating to the applicant, we calculate the distribution of optimal initial padding for those who apply. The mean padding level is 8%, with 70^{th} and 90^{th} percentiles equalling 18% and 31%, respectively. These results suggest that many applicants substantially exaggerate the true extent of their invention when they apply for patent rights.

We also compute weighted averages of padding as

$$\sum_{s} \frac{w_s}{\sum_{s'} w_{s'}} p_s,$$

where s denotes the application with padding p_s and our weights w_s are either the mean (over claims) of initial unpadded distances ($\bar{D}_s^*$) or initial unpadded values ($\bar{v}_s^*$). The weighted average of padding rises to 10% when weighted by values and 9% when weighted by distances, indicating that inventors increase padding for applications with claims that are more valuable and distant from the prior art (where such padding is less likely to induce the examiner to reject).

7.2 Examiner Parameters

Table 4 presents the estimates of the examiner parameters. To understand examiner costs and intrinsic motivation, we provide a slight digression on the units of examiner payoffs in the model, which we call "normalized credits."[34] The Office adjusts each examiner's credits based on their seniority and the technological complexity of applications. We use the same adjustments when we model payoffs for examiners.[35] These normalized credits are the unit of examiner payoffs; hence, we interpret intrinsic motivation costs, $\theta\mathcal{R}$, and delay costs per round π in terms of normalized credits.

We start by interpreting the parameters of intrinsic motivation. To our knowledge, these are the

[34] Appendix Section E.2 provides a detailed derivation of the examiners' credit structure.

[35] For example, an examiner receives two credits for granting a patent in the first negotiation round. We adjust these credits by dividing by a seniority factor (for example, by 1.25 for a senior GS-14 examiner) and multiplying by a technology correction (say, 29 for the relatively complex category of computer networks). Therefore, a GS-14 examiner in technology center "computer networks" receives 46.4 normalized credits for granting a patent in the first round. Tables E.1 and E.2 report the values of seniority and technology corrections across all seniorities and technology centers, respectively.

Table 4. Examiner Parameters Estimates

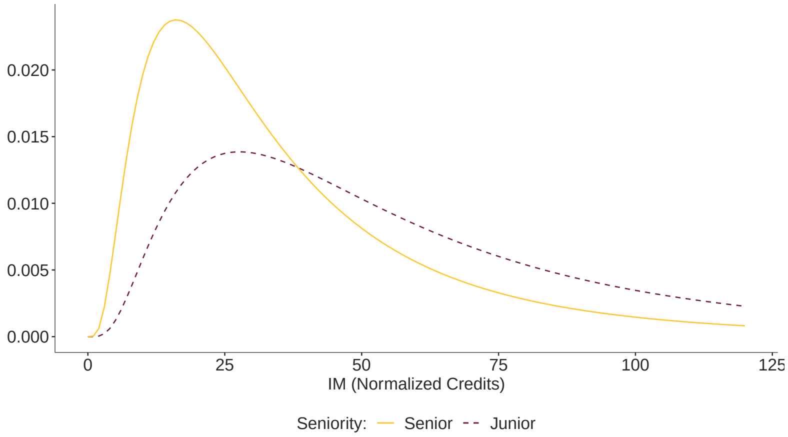
Parameter	**Symbol**	**Estimate**	**Standard Error**
Junior intrinsic motivation log-mean	$\mu_{\theta,\text{junior}}$	3.92	0.004
Senior intrinsic motivation log-mean	$\mu_{\theta,\text{senior}}$	3.38	0.005
Intrinsic motivation log-sigma	σ_θ	0.77	0.055
Delay cost log-mean	μ_π	0.19	0.006
Delay cost log-sigma	σ_π	0.27	0.015
Error standard deviation	σ_ε	0.02	0.000

Notes: This table provides the model parameters relating to the examiner. Standard errors are bootstrapped.

first structural estimates of intrinsic motivation in a public agency. We estimate σ_θ as 0.77, which implies, by the properties of the log-normal distribution, a coefficient of variation of 0.82 (82%). This estimate implies substantial variation in intrinsic motivation across examiners, even within seniority category. We estimate $\mu_{\theta,\text{junior}} = 3.92$ and $\mu_{\theta,\text{senior}} = 3.38$. These estimates imply that, on average, junior examiners are more intrinsically motivated than senior examiners. Figure 3 plots the distribution of intrinsic motivation for junior and senior examiners as implied by the log-normal assumptions. It is clear that the distribution of senior examiners' intrinsic motivation (yellow solid) is generally lower than that of junior examiners (maroon dashed). At least two countervailing forces influence the relationship between seniority and intrinsic motivation. Intrinsic motivation will fall with seniority if examiners become "jaded" with experience. However, selection cuts the other way since the least intrinsically motivated examiners are likelier to move to the private sector with higher remuneration. The evidence thus indicates that the jading effect dominates the selection effect.

To interpret the magnitude of intrinsic motivation, we calculate the associated cost for a median intrinsically motivated GS-12 (junior) examiner in a selected technology center 36 ("Miscellaneous" category). For this examiner, the seniority correction is one and the technology correction is 22.4. Recall that intrinsic motivation cost (in terms of normalized credits) is $C_{IM} = \theta\mathcal{R}$, where θ is the intrinsic motivation parameter, and R is the proportion of claims the examiner believes invalid. We divide C_{IM} by 22.4 to change the units back to pure credits. Hence, in terms of raw credits, this examiner's intrinsic motivation cost is $2.25R$, which means that the examiner faces a cost of 2.25 credits for knowingly granting a patent with *100%* of its claims as invalid. This cost is equivalent to the credits the examiner obtains for making *three* final rejections. This example is only an illustration, but our estimates generally imply that intrinsic motivation costs are sizeable relative to extrinsic rewards.

Figure 3. Density of examiner intrinsic motivation



Notes: Orange solid curve represents the distribution for senior examiners; maroon dashed curve represents the distribution for junior examiners. To interpret the x-axis, consider an examiner in technology center 36, where the technology correction is 22.4. Dividing the values on the x-axis by 22.4 yields the number of credits the examiner pays as an intrinsic motivation cost to a GS-12 examiner for granting a patent for an application on which every claim is invalid.

Next, we consider examiner delay costs. The coefficient of variation of examiner costs is 0.08, ten times smaller than examiner intrinsic motivation. Moreover, delay costs are estimated to be small, with the median cost for a GS-12 (junior) examiner in technology center 36 paying an equivalent of 0.05 credits to go an extra round on this particular application. The fact that these costs are so small suggests that examiners are not pressured explicitly to finish applications fast and that the opportunity cost of devoting more time to this application relative to the next on their desk is small. This finding is intuitive since the most time-consuming activity for the examiner is their initial literature search. Hence, continuing to make decisions on an application they have already reviewed is less time-intensive than starting a new application (though also less compensated).

Finally, we discuss examiner error parameters. Recall that examiner errors are normally distributed, with an estimated standard deviation and mean equal to $\mu_\varepsilon = 1 + \frac{1}{\theta}$, where θ is the examiner's intrinsic motivation. The error that an examiner draws multiplies padded distances to create the examiner's distance assessment. Since junior examiners are more intrinsically motivated on average, the mean of the junior examiners' error distribution is closer to one. We estimate the standard deviation of examiner errors to be 0.02, indicating that errors are modest, typically within 4% of the examiner-specific mean.

Calculating Examiner Errors

We compute statistics on two kinds of examiner errors by simulating the model with our baseline estimates. The first error occurs when an examiner grants a patent with invalid claims. We refer to this as a "type 1" error. We calculate that this happens for 19% of grants, suggesting that while examiners are screening out some invalid patents, nearly one in five applications contain some claims that should not have been granted. The last statistic represents the "extensive" margin of this type of examiner error; we can also calculate an "intensive margin" error. Among all granted claims, 7% are invalid (compared to 83% of claims whose unpadded distance is below the threshold), implying that most invalid patents contain only a few invalid claims.

We also calculate the weighted errors (focusing on the intensive margin), where weights reflect the distance of the claim from the patentability threshold. Among simulations s, let S_G be the set granted and j represent a claim. We calculate the measure

$$\sum_{s \in S_G} \sum_{j} \frac{w_{sj}}{\sum_{s' \in S_G} \sum_{j'} w_{s'j'}} E_{1sj,\text{int}}, \tag{12}$$

where $E_{1sj,\text{int}}$ is equal to one if claim j on simulation s is invalid (has a distance below the threshold) and zero otherwise, and the weight $w_{sj} = |\tilde{D}_{sj} - \tau_s|$, where τ_s is the threshold relevant to simulation s. The idea is to put more weight on errors where claims are further away from the threshold (making the error more "egregious"). If the weighted average is lower than the unweighted average, it implies that errors occur in marginal cases in which it is not obvious whether the patent is valid. Indeed, the weighted error is 2%, much lower than the unweighted value of 7%, suggesting that most errors occur in cases of marginal validity.[36]

The other kind of "error" (or "undesirable" outcome) occurs when an applicant abandons an

[36] We can also calculate weighted averages for the extensive margin type 1 errors. Let E_{1s} be equal to one if application s contains at least one valid claim when granted, and zero otherwise. Then we calculate

$$\sum_{s \in S_g} \frac{w_s}{\sum_{s' \in S_g} w_{s'}} E_{1s},$$

where w_s is

$$w_s = \begin{cases} \frac{1}{\#\text{inv}_s} \sum_{j \in \text{inv}_s} \mid \tilde{D}_{sj} - \tau_s \mid & \text{if} \quad E_{1s} = 1 \\ \frac{1}{M_{0,s}} \sum_{j=1}^{M_{0,s}} \mid \tilde{D}_{sj} - \tau_s \mid & \text{if} \quad E_{1s} = 0 \end{cases}$$

and inv_s is the set of invalid claims on simulation s, and $M_{0,s}$ is the total number of independent claims on simulation s. The weighted error in this case is 5.1%, similarly suggesting that most errors occur in marginal cases.

application that contains valid claims. We refer to these as “type 2” errors. Approximately 36% of abandonments have at least one valid claim. Strictly speaking, these are not a mistake by the examiner since they should only grant patents to applications on which *all* claims are valid. At the intensive margin, among all claims the applicant abandons, 18% are valid.

As in Equation (12), we calculate the measure:

$$\sum_{s \in S_A} \sum_{j} \frac{w_{sj}}{\sum_{s' \in S_A} \sum_{j'} w_{s'j'}} E_{2sj,\text{int}}, \tag{13}$$

where S_A represents the set of abandoned simulations, $E_{2sj,\text{int}}$ is equal to one if claim j on simulation s is valid and equal to zero otherwise, and the weights are the same as before – the distance of claim j from the threshold τ_s. In this case, the weighted error falls to 6%, again implying that abandonments occur on marginally valid claims rather than clearly valid ones. When we compute the extensive margin weighted error, the proportion of abandoned applications with at least one valid claim is 11%.[37]

7.3 Model Fit

Figure A.1 displays the values of simulated moments computed at the estimates alongside their analogs in the data. As expected, we match most of the internal moments well, though there are two exceptions. The first is the proportion of fully renewed patents, which we overestimate. The other exception is the second-round grant rate. This moment is difficult to match with our model because examiners have incentives to wait until the third round and obtain RCE credits if they do not choose to grant in the first round. Since examiners have incentives and targets across applications on their desks (docket management), they are more likely to grant in the second round than our baseline model predicts.

The real test of model fit is how well we match moments that are not used in the estimation procedure. Figure A.2 displays simulated moments and data moments for excluded moments described in Appendix G. We use untargeted percentiles on granted distances in rounds 1-6, the mean of distances for rounds 4-6, and the means and percentiles of round one rejection rates across seniority categories. We match these moments well. We also do a good job matching untargeted moments on the number of rounds (pooled across junior and senior categories), such as skewness and kurtosis, which is expected since we match moments on grants and abandonments by round and seniority of the examiner.

[37] We also compute the weighted version of the extensive margin type 2 errors in an analogous way to type 1 errors, as described in footnote 36.

7.4 Robustness

We run a series of robustness checks on our baseline model. Table A.3 provides the results. First, we examine changes to how we define the distance threshold for patentability. In the baseline, we define each examiner's "revealed" threshold as the minimum distance they grant and then take the threshold as the maximum of those values over examiners. As a robustness check, we use the first and fifth percentile of distances granted for each examiner, as opposed to the minimum, which allows for measurement error in their personal threshold. The parameter estimates are generally robust, and our qualitative conclusions are unchanged.

Second, we change the discount factor from 0.95 in the baseline to 0.99. In this case, most parameters are robust, with the notable exception of the parameters of the distribution of examiner delay costs, which are higher when β is higher. This feature is not surprising since both the discount factor and delay costs help to explain the same phenomenon: examiners choosing not to extend the negotiation process. This point also applies to applicant fighting costs, though with less severity.

Finally, we broaden the definition of "senior" examiners to include GS-13 and GS-14 (the baseline is GS-14 only). In our baseline model, we found that intrinsic motivation is stronger for junior examiners than seniors, and we want to check whether this finding is robust to how we define senior examiners. The results show that this finding is robust: a broader definition of seniority *increases* the difference we find in the baseline model.

8 Counterfactual Analysis

We use the estimated model to conduct a series of counterfactual analyses to examine the impacts of various reforms on the speed and quality of the screening process and the degree of padding in patent applications. The counterfactual scenarios we examine include removing intrinsic motivation and changing the level of patent office fees, the number of allowable rounds in the process, and examiner extrinsic incentives (credits).

Table 5 presents the results. We focus on four endogenous outcomes. The first is the proportion of applicants who choose not to apply for a patent on their developed invention. The second is the applicant's choice of how much to pad the application. The third set of outcomes is the proportion of grants in round one and the average number of rounds (speed of resolution). The fourth set, relating to screening quality, is the proportion of granted patents with at least one invalid claim (type 1 error at the extensive margin) and abandoned applications with at least one valid claim (type 2 error at the extensive margin). We note but do not report that the changes

Table 5. Counterfactual Experiments

Counterfactual	**Not Apply**	**Pad**	**Rounds**	**R1 Gr**	**T1**	**T2**
	(%)	(%)		(%)	(%)	(%)
Baseline	6.3	8.0	2.5	11.5	18.8	36.5
25K Round Fee	8.4	6.6	2.4	12.8	18.0	37.9
50K Round Fee	12.2	5.9	2.4	14.2	17.7	39.8
Three Rounds	27.2	3.6	2.1	15.1	15.9	49.2
Two Rounds	51.1	0.8	1.6	25.8	11.9	56.1
One Round	79.6	-2.3	1.0	98.4	0.5	91.7
15% IM	3.9	8.0	2.1	30.2	89.3	22.4
Credit↘	6.3	7.9	2.5	11.5	18.7	36.3
Credit↘ + 15% IM	3.4	18.1	2.1	32.8	88.9	17.7

Notes: "Not Apply' is the percent of inventors who do not apply for a patent; Pad is the mean level of padding. Rounds is the mean number of rounds. "R1 Gr" is the percent of applications granted in Round 1. T1 represents the proportion of granted patents with some invalid claims. T2 represents the proportion of abandoned applications with some valid claims.

in these errors at the intensive margin errors are similar.

Fees

In the baseline, there are relatively low fees for applying for a patent, finalizing the grant, and entering a request for continued examination. In the first counterfactual, we introduce a substantial per-round fee that the applicant must pay for *each* negotiation round (not just for an RCE).[38] This fee acts as a marginal cost per round of negotiation. Since each round is now more expensive, applicants have increased incentives to exit the patent process as soon as possible, and less incentive to apply in the first place. A substantial $50,000 fee for every extra round reduces padding by a quarter (from 8.0% to 5.9%) and slightly reduces the mean number of rounds,

[38] We also consider substantially increasing the *application* fees to as much as $50,000. However, because this is a fixed fee paid upon application, provided it is still profitable to apply, applicants will not change their padding decision. Even at this level, the fee does not materially alter average padding and, since there is practically no change to the proportion of inventors who choose to apply, introducing an application fee acts mainly as a transfer from applicants to the Patent Office, with minimal changes to quality or speed of prosecution. If the additional resources from the higher application fee were reinvested in patent office examination, there would be improvements. This finding—that application fees only really help if they are reinvested—is similar to the findings in Schankerman and Schuett (2022), who use a completely different theoretical model and data.

from 2.5 to 2.4. The proportion of grants in round one increases from 11.5% to 14.2%, reflecting the reduced padding and the fraction of granted patents with some invalid claims (type 1 error) falls slightly. However, the rounds fee increases type 2 error – rising from 36.5% to 39.8%. The trade-off between these two types of errors is a feature of many of the counterfactuals we analyze.

It is at first surprising that per-round fees as high as $50,000 do not substantially change the speed or quality of patent prosecution. The explanation is that the private value of patent rights is large enough to make applying for a patent on many of these inventions worthwhile, even with high per-round fees. Fees would have to be much higher to substantially impact outcomes.[39]

Restricting the Number of Rounds

Instead of using fees, we consider limiting the maximum number of rounds of negotiation between the applicant and examiner. We consider a maximum of three rounds, then a maximum of two rounds (equivalent to removing all RCEs, allowing only one round of interaction between applicant and examiner), and finally, we allow for only one round (that is, no negotiation between applicant and examiner so that the examiner's decision is final). These counterfactuals are motivated by a 2007 U.S.PTO proposal to restrict the number of RCEs. The proposed rule-making was challenged in federal court, which judged the restrictions as an overreach of Patent Office authority.[40] The court decision did not consider the quantitative impact of such changes on patent office screening quality or its welfare effects, which our paper makes possible.

Round restrictions have material consequences on screening outcomes. Removing all RCEs (allowing only two rounds) would lead to half of all inventors not applying for a patent and would virtually eliminate padding. In this case, 25.8% of applicants are granted in the first round, and because applicants respond to the restriction by reducing padding, the proportion of patents granted with invalid claims falls. In particular, with only one opportunity for negotiation, type 1 error falls sharply, from 18.8% to 11.9%.

[39] Of course, these fees would be significant for small firms or single inventors who may be cash constrained. However, round fees for small and micro entities could be reduced, as the Patent Office already does for other types of fees.

[40] The proposed changes are in *U.S.PTO Changes to Practice for Continued Examination Filings, Patent Applications Containing Patentably Indistinct Claims, and Examination of Claims in Patent Applications—the "New Rules"* (SmithKline Beecham Corp. v. Dudas, 541 F. Supp. 2d 805, 2008). The court decided that the "New Rules" were substantive and that the Patent Office did not have the rulemaking authority to make substantive changes, though the Court noted that the Patent Office could make procedural changes, such as fees. As we will show, one can achieve the same equilibrium number of rounds with an "equivalent" fee, so from an economic point of view, this distinction is problematic.

The disadvantage of limiting the scope for negotiation is that it increases the proportion of abandoned applications with valid claims. With no RCEs, this proportion rises from 36.5% to 56.1%. As with fees, making the process tougher for applicants through fewer allowable rounds generally reduces the granting of invalid claims and speeds up the process but leads to the abandonment of valid claims. As we discuss in the next section, granting invalid claims and not granting valid claims each imposes social costs, and we need to measure these to evaluate the overall impact of the reforms.

Finally, we compare the effectiveness of fees and round restrictions ("price versus quantity" instruments) by computing the equivalent per-round fee—in the sense of equalizing the mean number of rounds in equilibrium—to restrictions on the number of RCEs. The simulations show that the fee equivalent to removing all RCEs is a massive $600,000 per round. Using fees generally produces lower type 1 and type 2 errors than their rounds equivalents, but such fee levels are politically unpalatable.

Removing Intrinsic Motivation

Next, we evaluate the impact of removing intrinsic motivation by reducing it for every examiner to 15% of its original value.[41] Knowing that examiners will be more unwilling to grant invalid patents, only 3.9% of inventors do not apply, the number of rounds falls from 2.5 to 2.1, and the proportion of applications granted in round one almost triples, increasing from 11.5 to 30.2. Not surprisingly, type 1 error jumps sharply to 89.3%, while type 2 error declines. This counterfactual highlights the quantitative importance of intrinsic motivation on the quality of patent screening and confirms its potential salience for economic analyses of other public agencies.[42]

Removing Credits

Finally, we consider changes to the structure of credits for examiners. We remove all credits for the examiner after the first round. If it is the case that examiner costs of delay represent the marginal cost of an extra examination round, then such a policy change could be justified on efficiency grounds of "marginal cost pricing" since we estimated examiner delay costs to be

[41] We cannot fully remove intrinsic motivation because our specification of mean error being inversely related to IM implies that IM cannot be exactly zero. We provide the reason behind our choice of 15% in Section 9.

[42] Interestingly, increasing intrinsic motivation (not reported) does not have much impact in reducing padding or type-1 error. The explanation for this outcome is that examiners are already sufficiently intrinsically motivated to get most of the benefits, so further increases do not have much bite.

small.[43]

When we remove all credits after the first round in the baseline model, there are minimal, if any, impacts on any of the outcome variables. This result suggests that our baseline estimates of intrinsic motivation are sufficiently large for examiners to want to avoid granting invalid patents even in a context where they will receive no further extrinsic reward if they do so. This striking finding reflects the extent to which patent office examiners are intrinsically motivated. The results are not consistent with extrinsic incentives crowding out intrinsic incentives.

To complement this exercise, we also analyze the effect of removing all credits after the first round alongside reducing intrinsic motivation to 15% of its value (at any higher value of intrinsic motivation, removing credits has no material effect). In this case, we find non-trivial impacts of credits consistent with economic intuition. First, padding doubles, up to 18% (relative to 8.0% when only intrinsic motivation is changed) and first-round grants increase from 30.2% to 43.5%. Type 2 error declines because the increased padding means that abandonments are less likely to include valid claims.

These results indicate that extrinsic incentives and intrinsic motivation are *substitutes*, not complements, as sometimes found in the experimental literature (see Section 2 for citations): credits only work as an effective device to incentivize examiners when examiners are not intrinsically motivated (and even then, as we show in the next section, credits do *not* reduce social costs of screening).

In summary, these counterfactual experiments show that no reform we consider unambiguously improves both prosecution speed and quality. There is typically a trade-off: policies that make prosecution stricter lead to fewer grants of invalid patents but increased abandonments of valid applications. Evaluating reforms requires converting these outcomes into social costs, which we do in the following section.

[43] This counterfactual has limitations that the others do not because our model is focused on optimal decisions on a given patent application. It does not incorporate any interactions between different applications the examiner faces, such as optimizing docket management across applications (including meeting quarterly or annual targets). This counterfactual is best thought of informing an examiner that for one of the new applications in their docket, they will only receive credits for the first round.

9 Quantifying the Net Social Costs of Patent Screening

We classify net social costs into three categories: *type 1*, *type 2*, and *prosecution* costs. Type 1 costs refer to the costs induced by granting invalid claims. Type 2 costs refer to the social value of inventions that are not developed ex ante because of the potential threat of not being granted valid claims. Type 1 benefits refer to the social value of inventions that would not be developed ex ante without type 1 error. Type 2 benefits refer to the ex post deadweight loss *not* incurred when inventors abandon valid claims. Prosecution costs are the Patent Office's costs of examining applications plus the legal fees incurred by the applicants during the prosecution process. In what follows, we summarize our quantification approach; full details are in Appendix H. We start with the costs of each type of error and then discuss the benefits, with their difference defining net costs.

9.1 Type 1 Costs

There are two sources of costs from type 1 error: the deadweight loss associated with the royalties extracted by the patentee and the litigation costs associated with legal challenges against invalid patents that are granted (and that are valuable enough to warrant a challenge).

Deadweight loss from royalties

We assume that the patentee charges the Arrow royalty equal to the unit cost savings due to the invention, Δc. The deadweight loss from royalties depends on the market structure of licensees. Our baseline specification is perfect competition among licensees, with a linear demand and constant unit cost.[44] In this case, the deadweight loss is

$$DWL = \frac{1}{2}\Delta\wp\Delta q = \frac{\lambda}{2}\frac{\Delta\wp}{\wp}\tilde{V},$$

where $\wp$ is the initial price (without the royalty associated with the claim), $\Delta\wp = \Delta c$ with perfect competition, $\tilde{V} = q\Delta\wp$ denotes total *royalty payments*, and λ is the elasticity of product demand (in absolute value).[45] To calibrate this expression, we follow Schankerman and Schuett (2022), who estimate the ratio of corporate licensing revenue from intangible industrial property to R&D at 39.3%. Multiplying this ratio by the ratio of R&D to sales in manufacturing in 2002 (4.1%), we take $\frac{\Delta\wp}{\wp} = 1.61\%$. We do the computation for values of the demand elasticity $\lambda \in (1, 3)$ and

[44] In Appendix H, we extend the approach to Cournot competition. Our calibration indicates that this extension yields quantitatively very similar results.

[45] For invalid patents, we cannot use the model estimates of values of patent rights V to represent royalty payments for invalid patents $\tilde{V}$, since our estimates of V are contaminated with potential legal costs (explained in the next subsection). In Appendix H, we explain how we overcome this challenge to calculate type 1 costs.

report $\lambda = 2$ in the main analysis (qualitative conclusions hold for the other values).

Cost of litigation on invalid patents

The social cost of type 1 error also involves litigation costs on invalid patents. Not all invalid patents are "exposed" to litigation because their private value is not large enough to justify the litigation expense. Letting $G_{\tilde{V}}(\cdot)$ denote the distribution of the value at stake $\tilde{V}$, we take the proportion of patents not exposed to litigation from Schankerman and Schuett (2022) ($\check{v} = 89.6\%$) and calculate the $\check{v}^{st}$ percentile of the value at stake distribution, $\check{V} = G_{\tilde{V}}^{-1}(\check{v})$. Then, all patents with $\tilde{V}$ exceeding the threshold $\check{V}$ are exposed to litigation.

The social cost for invalid patents not exposed to litigation is only the deadweight loss from royalties. From Schankerman and Schuett (2022), exposed invalid patents have a 16.3% probability of being litigated, in which case, we assume that courts are perfect and thus always invalidate wrongly granted claims. In this case, the social cost is the sum of litigation costs for the patentee and challenger, each denoted $\mathcal{C}(\tilde{V})$.[46] The remaining 83.7% of exposed invalid patents are not litigated and only impose the deadweight loss.[47]

In summary, the expected social cost of granting an invalid patent of value $\tilde{V}_s$ is

$$S_{1s} = I_s DWL_s + (1 - I_s)\left[0.837 \cdot DWL_s + 0.163 \cdot 2\mathcal{C}(\tilde{V}_s)\right], \tag{14}$$

where $I_s = 1(\tilde{V}_s \leq \check{V})$ is an indicator equal to one if the patent is not exposed to litigation. Then, the total type 1 cost is

$$T_1 = \sum_{s \in S_G} E_{1s} S_{1s} \tag{15}$$

where E_{1s} is equal to one if a granted application $s \in S_G$ is invalid and zero otherwise.

9.2 Type 2 Costs

From the ex post perspective, there is *no social cost* from type 2 errors because the innovation has already been produced and the R&D cost is sunk (this is essentially ex post hold-up). Therefore, it only makes sense to analyze the social cost of type 2 errors from the ex ante (incentive) perspective. Type 2 error reduces the expected value of patent protection for the inventor and,

[46] We take $\mathcal{C}(\tilde{V})$ as linear in $\tilde{V}$ and calibrate the coefficients using AIPLA data.

[47] Patentees with invalid patents can pre-empt a challenge by charging a royalty payment (typically a lump sum) equal to the cost of litigation for the challenger (this is commonly referred to as "trolling" behavior). For these cases, the social cost is only the deadweight loss associated with the patent, since the payment is a pure transfer from the licensee to the patentee (we ignore possible R&D incentive effects of the transfer). See Schankerman and Schuett (2022) for more discussion.

thus, the ex ante decision of inventors to develop their (exogenous) ideas. We want to calculate the social value of the set of socially valuable inventions that are *not* developed when there is the possibility of type 2 error but which *would be* developed in the absence of type 2 error. This task requires us to construct a simple model of development. We emphasize that we do not require this extension to estimate the screening model.

The decision to develop an idea into an invention depends on three things: the ex ante value of patent rights (Γ^*), the value of the invention without patent rights (π), and the development cost (κ). To compute Γ^*, we use our model to calculate the ex ante value of patent rights (net of all costs), as in Equation (1). To calculate the private value of the invention without patent rights, we define the patent premium (ξ) as the percentage increase in private value due to patent protection. Hence, for positive Γ^*, by definition $\Gamma^* = \xi\pi$, implying a set of values of π. We assume that the patent premium is constant across inventions and calibrate it based on existing estimates from the literature on patent renewal models (Schankerman, 1998).[48] For the cost of developing an idea into an invention, κ, we draw values from the distribution estimated by Schankerman and Schuett (2022).[49]

An inventor *does not* invest to develop an idea i if

$$ND_i \equiv \mathcal{B}_i - \kappa_i \leq 0,$$

where $\mathcal{B}_i \equiv \pi_i + \max\{\Gamma_i^*, 0\}$ is the private benefit of development. An idea is socially valuable to develop if the net social benefit of development,

$$S_{2i} \equiv \frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}\mathcal{B}_i - \kappa_i,$$

is positive (where ρ_{priv} and ρ_{soc} denote the private and social rates of return). We use a conservative estimate of $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}} = 2$ from Bloom, Schankerman, and Van Reenen (2013).

Let Υ_0 denote the set of ideas that are socially beneficial to develop ($S_{2i} > 0$) but which are not developed ($ND_i \leq 0$). To calculate type 2 social cost, we compute the subset of Υ_0, which we denote Υ_1, that *would* develop in the absence of type 2 error. To do this, we simulate the

[48] This is a strong assumption, but it is not feasible to identify π_s if we allow the patent premium to vary. The reason is that we do not have any information on who develops their ideas, which might allow us to back out π from the decision to develop and our estimated value of Γ^*. Furthermore, we must specify π for inventions with negative ex ante value of patent rights. To do this, we draw from the distribution of π created from positive values of patent rights.

[49] An alternative approach is to assume that inventors do not know their development cost, and thus use the mean cost $\bar{\kappa}$. We experimented with this alternative and the values of type 2 social cost are similar in magnitude.

outcome from a "counterfactual" patent prosecution where, at the point of patent abandonment, the inventor obtains the value of all valid claims in that patent. By definition, in this scenario, all abandoned claims are invalid, so there is no type 2 error. Let Γ' denote the expected value of patent rights in this new scenario. The idea i would be developed in this scenario if

$$ND_i' \equiv \pi_i + \max\{\Gamma_i', 0\} - \kappa_i > 0.$$

We then compute type 2 costs as

$$T_2 = \sum_{i \in \Upsilon_1} S_{2i}, \tag{16}$$

where Υ_1 is the subset of Υ_0 with $ND_i' > 0$. This is the set of ideas that are socially beneficial to develop that are not developed in the scenario with type 2 error, but that would be developed in the absence of any type 2 error.

9.3 Patent Prosecution Costs

The social cost of patent prosecution for each application s consists of two components: applicant legal costs of amending the application each round and Patent Office administrative costs. The amendment cost is the per-negotiation cost $F_{\text{amend},s}$ drawn from the estimated distribution, multiplied by the equilibrium number of negotiations for application s (equal to the number of rounds r_s minus 1). For the administrative cost, we calculate the patent operations budget per application as \$4,117 (in 2018 dollars). This value excludes patent office fees, as these are transfers from the applicant to the patent office, as well as loss in patent value associated with pre-grant obsolescence since that, too, is a transfer from the applicant to the owner of the invention that superseded it. We divide this by the average number of rounds across all simulations and by the average number of independent claims in an application to create the average patent office cost per round and claim, denoted by RCC. Then, the total social cost of patent prosecution is

$$T_3 = \underbrace{\sum_s (r_s - 1) F_{\text{amend},s}}_{\text{Applicant Fighting Costs}} + \underbrace{\sum_s M_{0,s} r_s RCC}_{\text{Office Costs}}, \tag{17}$$

where $M_{0,s}$ is the initial number of claims in application s.

9.4 Benefits of Type 1 and Type 2 Errors

There are also benefits from errors. In the type 1 case, when invalid patents are incorrectly granted, the ex ante incentives for inventors to develop and patent their ideas are increased. This is analogous to the *costs* of type *2* error. We compute these benefits as the sum of social development benefits from welfare-enhancing projects that would not be developed without type

1 error but that are developed with type 1 error.[50] The method is similar to the approach described in Section 9.2.

Further, there are benefits from type 2 errors. Not granting valid patents saves the deadweight loss on those patents. We compute these benefits as described in Section 9.1. Note that there is no benefit associated with litigation cost savings since, under our assumption of costly but perfect courts (always upholding valid patents and overturning invalid ones), valid patents that are granted would not be challenged.

To see this, suppose the threshold is too low (the conventional wisdom) so that some patents are considered "valid" and are granted despite not being welfare-enhancing. We would not count these as a type 1 error, and they would not contribute to type 1 social costs. Hence, we would understate type 1 costs (and type 2 benefits). By a similar argument, we would overstate type 2 costs and type 1 benefits, so we would understate net type 1 social costs and overstate net type 2 social costs. The reverse would be true if the threshold used by the Patent Office were too high. We are unable in this paper to test whether the threshold is socially optimal. It is an open research question what the "optimal" distance threshold would be that corresponds to the socially optimal rule, which is to grant patents only to inventions that are welfare-enhancing and not otherwise developed (Schankerman and Schuett, 2022).

One important point to note is that the quantification of net social costs in this section is based on the presumption that the patentability threshold used by the Patent Office corresponds to the social optimum, that is, the threshold that only grants patents to inventions that are welfare-enhancing but would not be developed without patent rights. To see this, suppose the threshold is too low (the conventional wisdom) so that some patents are considered "valid" and granted despite not being welfare-enhancing. We would incorrectly not count these as a type 1 error, so they would not contribute to our measure of type 1 social costs. Thus, we would understate type 1 costs (and type 2 benefits). By an analogous argument, we would overstate type 2 costs (and type 1 benefits). Therefore, if the threshold is too low, the consequence is that we would understate net type 1 social costs and overstate net type 2 social costs. The reverse would be true if the threshold used by the Patent Office were too high. We are unable in this paper to test whether the threshold is socially optimal. It is an open research question to determine the "optimal" distance threshold, that is, the one that grants patents only to inventions that are welfare-enhancing and not otherwise developed (Schankerman and Schuett, 2022).

[50] The "counterfactual" patent prosecution in this case is one where, at the point of patent grant, the inventor only obtains the value of the *valid* claims in the patent.

9.5 Social Costs in Counterfactual Reforms

Table 6 summarizes the three components of net social costs for the baseline model and the set of counterfactual reforms.[51] The baseline row approximates the net social costs associated with a yearly cohort of ideas, averaged over 2011-2013 (Appendix H explains how we calibrate the annual number of ideas). Subsequent rows provide the net social costs in that counterfactual scenario. All values are adjusted for inflation, presented in 2023 U.S. dollars.

In the baseline, total type 1 net costs equal \$6.4bn, total type 2 net costs are \$1.5bn, and prosecution costs equate to \$17.6bn. In the final column, we sum these three net costs and estimate the total net social cost of patent screening at \$25.5bn. This total constitutes 6.5% of total R&D performed by business enterprises in the U.S. in 2011.[52]

Introducing a per-round fee lowers type 1 and prosecution costs because it discourages applications and lowers padding for those that do apply. This, in turn, implies that fewer grants are invalid and that grants occur in fewer rounds. However, a round fee increases type 2 costs as applicants are more likely to abandon with some valid claims in a scenario with high negotiation fees. With a \$25,000 round fee, the latter effect dominates, so the total net social cost increases by a very modest 1.9%. As mentioned earlier, for sufficiently large rounds fees (likely to be politically infeasible), the reductions in type 1 and prosecution costs eventually dominate. Further, in these counterfactuals, the extra revenues generated by the fees are *not* reinvested in more intensive or faster examinations. If they were reinvested, social costs from introducing fees would be mitigated or even converted to social gains.

Restrictions on the allowable number of negotiation rounds have qualitatively similar effects on social costs as rounds fees, but the impacts are much larger. Removing all RCEs (two rounds) yields a 10.4% fall in total social costs relative to the baseline. Restricting the process to one round reduces net social costs by 45%.

Removing intrinsic motivation (down to 15% of its original level) increases the total social cost by

[51] The table presents the values of net social costs for $\lambda = 2$, $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}} = 2$, $\xi = 0.1$, and development costs drawn. The qualitative conclusions are similar for a range of other parameter values. In Appendix A, we provide results for the cases of a 5% patent premium with $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}$ equal to 1.5 and 2, and a 10% patent premium with $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}$ equal to 1.5. We do not present results for different values of lambda because quantitative values in this case are very similar to the baseline.

[52] It is worth noting that this is at the lower end of estimates of the private value of patent rights (Pakes, 1986; Schankerman, 1998). This suggests that the patent system, as it is currently configured, generates net positive social value. For similar findings in a different framework, see Schankerman and Schuett (2022).

Table 6. Net Social Costs of Patent Prosecution

Counterfactual	$\boldsymbol{T_1}$	$\boldsymbol{T_2}$	$\boldsymbol{T_3}$	**Total**
Baseline ($Bn)	6.4	1.5	17.6	25.5
25K Round Fee	5.9	3.7	16.4	26.0
50K Round Fee	6.1	6.3	15.1	27.5
Three Rounds	4.9	10.1	10.2	25.1
Two Rounds	2.9	15.6	4.7	23.1
One Round	0.0	13.4	0.7	14.1
15% IM	25.8	2.3	15.0	43.0
Credit$\searrow$	6.4	1.5	17.6	25.5
Credit$\searrow$ + 15% IM	15.9	4.0	15.8	35.7

Notes: Equation (15) defines T_1; Equation (16) defines T_2, respectively; Equation (17) defines T_3; Total sums the three kinds of costs. The "baseline" row provides the total social costs in billions of 2018 U.S. dollars.

68.6%. When examiners have almost no intrinsic motivation, they are willing to grant applications fast, even if they are padded. As a result, administrative costs fall when intrinsic motivation is removed.[53] However, the grants of patents with invalid claims cause type 1 net costs to triple and consequently lead to an overall rise in net social costs. This finding confirms the critical role that intrinsic motivation plays in this public agency.

Finally, with the baseline level of intrinsic motivation, removing all examiner credits after the first round for one examination has almost no effect on social costs – precisely as we would expect, given the negligible changes to any endogenous variables. In fact, examiners' intrinsic motivation must be as low as 15% of original values for credits to have any effect on net social costs. With 15% intrinsic motivation, type 2 gross (and net) costs, prosecution costs, and type 1 *gross* costs all increase when credits are removed. As a result, when intrinsic motivation is lowered by 85%, removing credits increases total *gross* social costs. Yet, total *net* social costs *decrease*, suggesting that credits are counter-productive even when intrinsic motivation is low. This finding is driven by a large increase in type 1 *benefits* (and hence a decrease in type 1 *net* social costs) from removing credits. This result highlights the importance of accounting for the

[53] The decrease in prosecution costs is countervailed by the fact that when intrinsic motivation is low, there is an extensive margin increase in the number of inventors applying for patent rights.

increased development from relaxed patent granting, as opposed to just the ex post social costs that arise through deadweight losses and litigation.

10 Conclusion

In this paper, we develop and estimate a structural model of the patent screening process. The model incorporates incentives, intrinsic motivation, and multi-round negotiation between the examiner and applicant. The paper shows how structural modeling of the incentives and organization of innovation-supporting public agencies can be used to design reforms to improve agency performance. Our paper highlights the fact that, to analyze the impact of reforms on the effectiveness of screening, it is critical to incorporate *both* the agency's decision-making and the endogenous responses of applicants being screened.

Our key empirical findings are as follows. First, contrary to pervasive criticisms in the public debate, we find that patent screening is relatively effective, *given* the statutory and judicial standards for patentability within which the Patent Office is required to operate. Even though most patent applications do not initially meet the patentability threshold (in part because applicants strategically exaggerate the scope of their property rights), the examination process narrows these patents sufficiently to ensure that a much smaller proportion of granted patents are invalid. This fact implies that the high percentage of patent applications granted is a misleading indicator of the effectiveness of patent screening – the percentage of initially claimed property rights that are eventually granted is much lower.

Second, the effectiveness of patent screening is primarily driven by a high degree of intrinsic motivation among patent examiners. This result highlights the potential importance of intrinsic motivation in the context of other public agencies. In the presence of such high motivation, we find that extrinsic incentives are largely ineffective. However, extrinsic incentives do materially affect agency performance in a scenario where intrinsic motivation is virtually absent.

Third, our counterfactual analysis reveals that restrictions on the number of allowable rounds of negotiation (currently *absent* in the U.S. patent system) reduce the social costs of screening. This outcome can be replicated through an equivalent round fee for the applicant, but the required fees are too high to be politically feasible.

Finally, we estimate the total net social cost of patent screening at \$25.5bn per annual cohort of applications. This figure represents 6.5% of R&D in the United States performed by business enterprises. These costs include the administrative cost to the patent office, the transaction cost for applicants, the ex post cost of granting patents that do not meet the standard, and the ex ante (incentive) costs of not granting patents that do meet the standard.

This paper studies patent screening and instruments to improve its effectiveness at the *pooled* technology level. A fruitful extension would be to estimate the model for individual technology fields, such as biotechnology and software, which would allow for the evaluation of the differential effectiveness of various instruments in different areas. More generally, we hope this paper illustrates the value of using structural models to inform decisions on how to reform public agencies, particularly those that affect the allocation of R&D resources, including leading institutions like the National Institutes of Health and National Science Foundation, and similar institutions in other countries.

Appendices

Online appendices are available **here**.

References

Adda, J. and M. Ottaviani (2023): "Grantmaking, Grading on a Curve, and the Paradox of Relative Evaluation in Nonmarkets," *forthcoming in The Quarterly Journal of Economics.*

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

Ashraf, N., O. Bandiera, E. Davenport, and S. S. Lee (2020): "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services," *American Economic Review*, 110, 1355–94.

Ashraf, N., O. Bandiera, and K. Jack (2014): "No margin, no mission? A field experiment on incentives for public service delivery," *Journal of Public Economics*, 120, 1 – 17.

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2018): "Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules," *The Review of Economic Studies*, 86, 117–152.

Benabou, R. and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *The Review of Economic Studies*, 70, 489–520.

——— (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

Besley, T. and M. Ghatak (2005): "Competition and Incentives with Motivated Agents," *American Economic Review*, 95, 616–636.

Bessen, J. (2008): "The value of U.S. patents by owner and patent characteristics," *Research Policy*, 37, 932–945.

Bloom, N., M. Schankerman, and J. Van Reenen (2013): "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, 81, 1347–1393.

Cockburn, I., S. Kortum, and S. Stern (2003): *Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes*, Washington, DC: The National Academies Press.

Egan, M. L., G. Matvos, and A. Seru (2018): "Arbitration with Uninformed Consumers," *Working Paper.*

Federal Trade Commission (2011): *The Evolving IP Marketplace: Aligning Patent Notice and Remedies with Competition*, Washington D.C.: Government Printing Office.

Feng, J. and X. Jaravel (2019): “Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation,” *American Economic Journal, Applied Economics.*

Foit, L. (2018): “Understanding the USPTO Examiner Production System,” *Midwest IP Institute.*

Frakes, M. D. and M. F. Wasserman (2017): “Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data,” *The Review of Economics and Statistics*, 99, 550–563.

Galasso, A. and M. Schankerman (2015): “ Patents and Cumulative Innovation: Causal Evidence from the Courts,” *The Quarterly Journal of Economics*, 130, 317–369.

——— (2018): “Patent rights, innovation, and firm exit,” *The RAND Journal of Economics*, 49, 64–86.

Gaule, P. (2018): “Patents and the Success of Venture-Capital Backed Startups: Using Examiner Assignment to Estimate Causal Effects,” *The Journal of Industrial Economics*, 66, 350–376.

Gowrisankaran, G., A. Nevo, and R. Town (2015): “Mergers When Prices Are Negotiated: Evidence from the Hospital Industry,” *American Economic Review*, 105, 172–203.

Graham, S., A. Marco, and R. Miller (2018): “The USPTO Patent Examination Research Dataset: A window on patent processing,” *Journal of Economics and Management Strategy*, 27, 554–578.

Grennan, M. (2013): “Price Discrimination and Bargaining: Empirical Evidence from Medical Devices,” *American Economic Review*, 103, 145–77.

Hall, B. and J. Lerner (2010): *The Financing of R&D and Innovation*, vol. 1, Elsevier.

Jaffe, A. and J. Lerner (2004): *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It*, Princeton University Press.

Jalali, M., H. Rahmandad, and H. Ghoddusi (2015): *Using the method of simulated moments for system identification*, MIT Press.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): “Measuring Technological Innovation over the Long Run,” *American Economic Review: Insights*, 3, 303–20.

LANJOUW, J. O. (1998): "Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value," *The Review of Economic Studies*, 65, 671–710.

LE, Q. AND T. MIKOLOV (2014): "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 1188–1196.

LEMLEY, M. A. AND B. SAMPAT (2012): "Examiner Characteristics and Patent Office Outcomes," *The Review of Economics and Statistics*, 94, 817–827.

LI, D. (2017): "Expertise versus Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 9, 60–92.

LI, D. AND L. AGHA (2015): "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science*, 348, 434–438.

LU, Q., A. F. MYERS, AND S. BELIVEAU (2017): " USPTO Patent Prosecution Research Data: Unlocking Office Action Traits," *USPTO Economic Working Paper*.

MERGES, R. AND J. DUFFY (2002): "Patent Law and Policy: Cases and Materials," *Newark, NJ: LexisNexis*.

MERGES, R. P. AND R. R. NELSON (1990): "On the Complex Economics of Patent Scope," *Columbia Law Review*, 90, 839–916.

NELSEN, R. B. (2007): *An introduction to copulas*, Springer Science & Business Media.

PAKES, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755–784.

PRENDERGAST, C. (2007): "The Motivation and Bias of Bureaucrats," *American Economic Review*, 97, 180–196.

SAMPAT, B. AND H. L. WILLIAMS (2019): "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, 109, 203–36.

SCHANKERMAN, M. (1998): "How Valuable is Patent Protection? Estimates by Technology Field," *The RAND Journal of Economics*, 29, 77–107.

SCHANKERMAN, M. AND A. PAKES (1986): "Estimates of the Value of Patent Rights in European Countries during the Post-1950 Period," *Economic Journal*, 96, 1052–1076.

SCHANKERMAN, M. AND F. SCHUETT (2022): "Patent Screening, Innovation, and Welfare," *The Review of Economic Studies*, 89, 2101–2148.

THE ECONOMIST (2015): "Time to fix patents," *The Economist Group*, August 8th-14th, 9.