Zaby, Alexandra Karin; de Rassenfosse, Gaétan

Conference Paper

The Economics of Patent Backlog

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Patents, No. E11-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Zaby, Alexandra Karin; de Rassenfosse, Gaétan (2016) : The Economics of Patent Backlog, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Patents, No. E11-V3, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/145673

The Economics of Patent Backlog*

Gaétan de Rassenfosse** and Alexandra K. Zaby§

July 2016

Abstract

Patent offices around the world face massive backlogs of applications, which threatens to slow down the pace of technological progress. However, economists lack analytical tools to tackle the issue. This paper provides a model of patent backlogs inspired by the literature on traffic congestion. Inventors are heterogeneous with respect to their desired duration of patent pendency. They can accelerate or slow down the duration of pendency by adapting their filing strategy. We use our model to discuss three policy responses: increasing examination capacity; introducing penalty fees; and altering the value of pending applications.

Keywords: congestion model, patent backlog, pendency, strategic filing
JEL Classification: O34, R48

*The authors are grateful to current and former chief economists and senior economic advisors at seven patent offices (CIPO, EPO, IP Australia, JPO, USPTO and WIPO) for having shared their expertise. We guaranteed anonymity to our respondents. Dietmar Harhoff, Francisco Ruiz-Aliseda and Ying Lei Toh provided valuable comments.

**Ecole polytechnique fédérale de Lausanne, College of Management of Technology, Lausanne, Switzerland. E-mail: gaetan.derassenfosse@epfl.ch

§University of Tuebingen, School of Business and Economics, Mohlstrae 36, D-72074 Tuebingen, Germany. E-mail: alexandra.zaby@uni-tuebingen.de

1 Introduction

The world's largest patent offices face massive backlogs of applications. At the end of 2013, more than two million patent applications were awaiting examination at the European Patent Office (EPO), the Japan Patent Office (JPO), and the US Patent and Trademark Office (USPTO) (IP5 Offices, 2015, p. 71). Lengthy delays defer the introduction of new products to the market, create uncertainty for competitors, and may ultimately slow down the pace of technological progress.

Although excessive patent pendency can be detrimental to welfare, it may have important private benefits. Some inventors actively seek to postpone the grant decision in order to hide their inventions from competitors as long as possible or to adjust the scope of the claimed invention as the technology evolves. It is no surprise that patent attorneys have deployed various techniques to modulate the duration of patent pendency. For example they may write claims that are excessive and unclear (Sperber, 1970; Popp, Juhl and Johnson, 2004) or file continuations of the original application (Graham, 2006) in order to prolong pendency duration. Such filing strategies place further strain on the patent office's processing capacity and aggravate the backlog.

The present paper proposes a theoretical study of patent backlogs. It analyzes how inventors adapt their filing strategy in anticipation of the backlog—acknowledging the fact that these two dimensions affect one another—in order to reach the duration of pendency that they desire. We draw a parallel between patent applications facing a backlog and commuters facing traffic congestion thereby following a handful of other scholars (Osenga, 2005; Sharon and Liu, 2007; Marco and Prieger, 2009). Yet, to the best of our knowledge, the present paper is the first to propose a full-fledged model based on the parallel and to use it to discuss possible policy responses to the backlog.

The theoretical framework adapts the dynamic bottleneck model first developed in the traffic congestion literature in urban economics (Vickrey, 1969; Arnott, de Palma and Lindsey, 1990, 1993; Small and Verhoef, 2007). This framework is particularly relevant because it explicitly models a bottleneck so that comparative statics give straightforward insights on the interplay between filing strategy and backlog. The present study clarifies how best to incentivize inventors in order to mitigate the negative welfare effects caused by patent backlogs. It fits into the rich literature on the optimal design of patent systems (see, for example, DeBrock, 1985; Matutes, Régibeau and Rockett, 1996; Gallini, 2002; Acemoglu and Akcigit, 2012).

In the context of this paper, the duration of a "trip" from A to B corresponds to the duration

of patent examination (that is, the time between application and final decision date), and "road capacity" corresponds to the patent office's examination capacity. A "queue" develops when the workload caused by all pending patent applications exceeds the patent office's capacity, which slows down the examination process. Inventors, who anticipate the effect that the backlog will have on the duration of pendency of their application, manipulate their application in order to minimize their costs of pendency. This situation is similar to commuters deciding to adapt their departure time in order to minimize the cost of their journey.

We use the model to provide novel insights on policy responses to the backlog. The analysis shows that a key parameter for policy purposes is the proportion of applicants seeking to delay the grant decision. In order to gauge the relevance of each possible policy response, we asked seven current and former chief economists at patent offices for their informed estimate of this parameter. Overall, they believe that more inventors wish to delay the grant decision than to hasten it. Under this condition, the analysis shows that increasing examination capacity has a limited positive welfare effect. Regarding the use of penalty fees, a second policy response, the analysis shows that a linear penalty fee that targets backlog-inducing characteristics of the patent application or the patent prosecution process (for example, the number of claims) is more efficient than a stepped penalty fee—provided that fees are high enough. Finally, the analysis shows that it may be worth increasing the cost of patent pendency, for example, by systematically publishing patent applications after 18 months. Scholars can build on our model to conduct a congestion-pricing analysis of additional policy responses, such as fast-tracking applications or deferring examination.

The rest of the paper is organized as follows. Section 2 presents background information on key aspects of patent pendency. Section 3 stylizes the patent examination process and Section 4 introduces the dynamic patent congestion model. Section 5 discusses welfare considerations and policy implications arising from the model. The last section offers conclusions.

2 Background

Existing empirical research has documented the heterogeneity of inventors with respect to desired prosecution time. The entrepreneur relying on her patent to raise capital (e.g., Conti, Thursby and Thursby, 2013) and the company needing its patent granted as soon as possible in order to request injunction against an infringer are classic examples illustrating the need for a short pendency. The textbook example of need for a long pendency is that of an early-stage invention not fully developed yet but submitted to the patent office to secure the priority right.

Delaying the grant decision gives the inventor more time to assess the market potential of the invention and adjust the claims as the uses of the technology become clearer. Berger, Blind and Thumm (2012) report that such behavior occurs in the context of standard-setting negotiations, where inventors amend their pending applications to achieve conformity with standards under development. The strong increase of divisional filings after the EPO had introduced restrictions on their timing in 2010 suggests that a substantial amount of patent applicants have a general interest in delaying the examination process of their applications (see Harhoff, 2016).

Legal scholars have repeatedly argued that patent attorneys have considerable latitude in drafting patent applications and that well-drafted applications—that is, applications that are well documented, make narrow claims, and use precise language—are examined faster than poorly drafted patents (e.g., Sperber, 1970; Popp, Juhl and Johnson, 2004; Harhoff and Wagner, 2009; Mabey, 2010; Koenen and Peitz, 2012). Relating interviews with patent examiners, Popp, Juhl and Johnson (2004, p. 35) explain that delays become considerably longer when additional communications are needed between the examiner and the applicant. They report that the likelihood of a communication occurring increases with the number of claims and with claims that are unclear or too broadly defined, a fact that "very experienced attorneys" are well aware of. Lazaridis and van Pottelsberghe de la Potterie (2007) quantify the impact of the number of claims on pendency duration at the EPO. They estimate that two additional claims are associated with an additional communication, prolonging the examination process by a year. Koenen and Peitz (2012) discuss the various ways in which applicants can prolong patent pendency.

Observers of the patent system generally agree that lengthy pendency is detrimental to welfare (Palangkaraya, Jensen and Webster, 2008; Mabey, 2010; Graham and Hancock, 2014). Patent pendency is associated with uncertainty about property rights, which may defer the introduction of new products to the market and distort rival firms' investment decisions. For example, Gans, Hsu and Stern (2008) show that the probability of achieving a licensing agreement significantly increases after the patent is issued—that is, once uncertainty about the scope of the patent has been resolved. In a similar vein, pending patents and long lags between patent application and grant vitiate technology transactions by demanding disclosure without supporting appropriability (de Rassenfosse, Palangkaraya and Webster, 2016). Excessive pendency also opens the door to so-called submarine patents. Inventors in the United States have the right to refuse publication of their patent applications prior to grant (but only if they do not seek international extension). Unpublished applications go unnoticed by competitors performing freedom-to-operate searches, giving them a false sense of operating in

a patent-free environment and putting them at risk of subsequent holdups (Reitzig, Henkel and Heath, 2007). Investigating the effects of pendency on start-up firms, Farre-Mensa, Hegde and Ljungqvist (2015) find that excessive pendency has a dramatically negative impact on entrepreneurial growth and success: they conclude that two years of additional pendency have the same effect as rejection of the patent application.

In a nutshell, inventors have heterogeneous pendency preferences, and patent attorneys have the ability to influence pendency duration, creating both incentives and opportunities to game the system. These strategic behaviors occur in a context characterized by a massive backlog of patent applications, which affects pendency duration in a broad manner. Given that excessive grant delays are detrimental to welfare, it is important to understand the relations between filing behavior and the backlog.

The next two sections present a model of patent backlogs adapted from the traffic congestion literature. Note that the model focuses exclusively on filing strategies that use examiner time and, therefore, affect the pool of patents awaiting examination. In other words, it considers strategies with an externality effect on the backlog.[1]

3 Stylizing the patent examination process

3.1 Patent cohort

The model focuses on a cohort of patent applications filed at date $t_f = 0$ and awaiting examination. The demand for patent examination is perfectly inelastic, in that inventors cannot withdraw their patent applications. Before date t_f the patent office is not congested and works at an exogenously given and constant examination capacity and quality—that is, it can neither hire more examiners nor reduce examination time per application. We will relax this assumption in Section 5 when considering policy responses to the backlog. But for the moment we state:

Assumption 1 *Examination capacity and quality are fixed.*

Inventors in the cohort are heterogeneous with respect to the desired duration of pendency t_p. Given application date $t_f = 0$, desired pendency duration corresponds to the preferred end date

[1]Not all filing strategies use examiner time. For example, inventors at some patent offices have the option to postpone the grant decision by deferring examination: the grant decision is delayed but no extra workload is imposed on examiners. Similarly, inventors can instruct their patent attorneys to delay their responses to examiners' communications.

of the prosecution process. We assume that desired pendency duration is uniformly distributed on the interval $[\underline{t}_p, \overline{t}_p]$:

Assumption 2 *Desired pendency duration is uniformly distributed.*

Note that the model is agnostic with respect to the grant outcome: some patent applications in the cohort will be granted whereas others will be refused. This fact is of little concern to the analysis. What matters from the patent office's point of view is that both granted and refused patent applications consume examiner time. The grant outcome would matter if it depended on the filing strategy (for example, if patent applications written in less precise ways were less likely to be granted). However, as Lemley and Shapiro (2005, p. 75) have emphasized, it is very difficult for the patent office to refuse a patent application even if the invention described is "broad and rather vague." More generally, we note that patent examiners do not assess the quality of the drafting style but the technological merit of a patent application. We therefore assume that the grant outcome is independent from the filing strategy and state:

Assumption 3 *The filing strategy does not affect the probability of grant.*

3.2 Workload of the patent office

We divide examination time into a discrete number of examination steps. The number of steps associated with a patent application depends on the filing strategy of the inventor: a low-workload application requires fewer examination steps than a high-workload application. The total workload of the patent office is simply the sum of examination steps associated with all pending applications. Without congestion the patent office is able to adhere to the "standard" time necessary to work through the examination steps of every application. If, for example, the patent office has the capacity of examining one step per time interval, pendency duration without congestion would simply correspond to the number of necessary examination steps. With congestion pendency duration increases beyond standard examination time as the number of examination steps of all patents awaiting examination exceeds the capacity of the patent office. We assume that the patent office is congested from $\underline{t} \geq t_f$ onward and state:

Assumption 4 *The patent office faces a backlog.*

This assumption implies that the examination time of a patent application is longer than its standard examination time, that is, all patent applications face "excess examination time." This situation is similar to that currently faced by patent applicants at the main patent offices

around the world. However, we will discuss how inventors can influence pendency duration with their filing strategy.

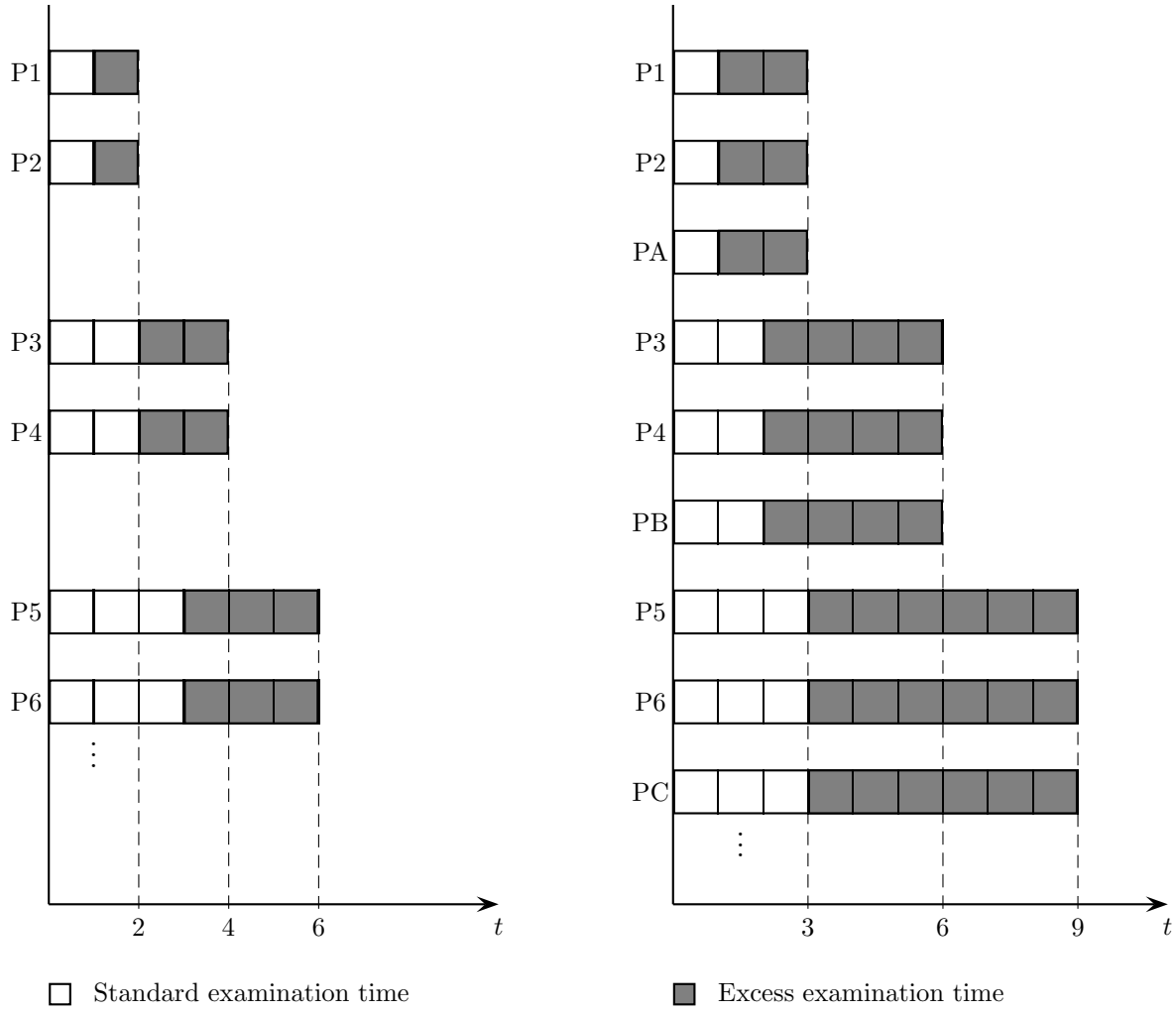
3.3 Patent backlog

The backlog develops because the workload exceeds the patent office's processing capacity such that the examination steps cannot be processed in the standard time. The extent of the backlog is given by the aggregate number of examination steps waiting to be processed, meaning that the backlog for the cohort monotonically decreases over time subject to the patent office's processing capacity.[2]

Figures 1(a) and 1(b) illustrate how the backlog affects pendency duration. In the case of unconstrained capacity at the patent office, examiners are able to process all patent applications in their standard examination time (white blocks). In Figure 1(a) patent applications P1 and P2 require one examination step, applications P3 and P4 two, and applications P5 and P6 three examination steps. Thus, these applications are processed within one (P1, P2), two (P3, P4), or three (P5, P6) time intervals if there is no congestion at the patent office. Given constrained examination capacity of one patent application per time interval, applications P1 and P2 face excess examination time after their standard examination time elapses at the end of the first time interval. Because examiners can now process only one half of the two applications P1 and P2 per time interval, these applications face excess examination time of one additional time interval. Applications P3 and P4 face two intervals of excess examination, and patents P5 and P6, with the higher workload of three examination steps, face three intervals of excess examination time.

Moving to Figure 1(b) it becomes clear that the flow of patent applications reaching their standard examination time is key to the severity of the backlog. Whereas in Figure 1(a) two applications per time interval reach their standard examination time, in Figure 1(b) three applications per time interval do so, which causes an increase of excess examination time for all applications. P1 and P2 now face two intervals of excess examination time; applications P3 and P4 face four intervals of excess examination time; and applications P5 and P6 now face six intervals of excess examination time. Thus, it is not the number of applications that matters but how time-consuming it is to examine them.

[2]The definition of backlog in this paper is thus the "excess" of applications over office capacity as discussed in Mitra-Kahn et al. (2013). The authors present other definitions and attempt to measure backlog in a consistent manner across offices.



(a) Two patents per time interval reach standard examination time

(b) Three patents per time interval reach standard examination time

Figure 1: Excess examination time due to backlog

3.4 Pendency costs and filing strategy

The extent of the costs associated with excess examination time (simply, "waiting costs") depends on the additional examination time caused by the backlog. Such waiting costs capture, for example, profits lost because of delayed royalties (Gans, Hsu and Stern, 2008). Another example is a situation in which a patent application by a start-up firm is issued after a debt-funding agreement has been reached, such that financing terms are less favorable than they might have been had the patent application been issued earlier (de Rassenfosse and Fischer, 2016).

By contrast, applicants also face costs associated with the deviation between actual pen-

dency and desired pendency (simply, "deviation costs"). Deviation costs are incurred when inventors manipulate their filing strategies. For instance, an inventor may decide to reduce the number of steps required to examine her patent application by drafting a narrower specification. By doing so, however, she reduces the scope of her patent and may make it weaker in case of litigation. Under other circumstances it may be optimal for her to increase the number of examination steps, which will entail higher drafting costs. Inventors can influence both cost types by the choice of their filing strategy.[3] The filing strategy affects the number of examination steps and consequently the standard time necessary to complete the examination process. Choosing a low-workload filing strategy reduces the number of examination steps, whereas choosing a high-workload filing strategy increases the number of examination steps. A high-workload application with a later examination date may thus—despite higher deviation costs—lead to a higher utility because it is accompanied by lower excess examination time: When a patent application reaches its standard examination time later, excess examination time is possibly decreasing again, meaning that it is lower as compared with a situation where the application reaches its standard examination time earlier. A low-workload application with an earlier examination date may—despite decreasing excess examination time—lead to a lower utility as it comes along with higher deviation costs.[4]

[3]It is useful to draw an analogy with traffic congestion to illustrate the waiting and deviation costs. The *waiting costs* capture the lost productivity that commuters incur while stuck in traffic. *Deviation costs* are costs associated with arriving too early or too late to work because commuters, anticipating congestion, leave earlier or later than desired in order to avoid being stuck in traffic. Take the example of a morning traffic situation in which a given number of commuters pass through a certain traffic light. All commuters want to arrive at work between 7 a.m. and 9 a.m. Starting before 7 a.m., more and more cars arrive at the traffic light, and a queue starts to develop. First, the queue grows over time, but then, as time passes, the number of cars arriving at the traffic light decreases. The queue then begins to dwindle, until at some time after 9 a.m. the last commuter passes the traffic light. In deciding when to leave for work a commuter trades off the cost of the time spent waiting in the queue against the cost of arriving at work earlier (later) than desired. If she leaves early, she faces a short queue, if any, but arrives too early at work. If she leaves closer to her desired arrival time she faces a long queue but arrives on time.

[4]It is worth noting that these endogenous filing strategy (i.e., scheduling) decisions together with the interplay of waiting and deviation costs are a clear demarcation between dynamic bottleneck models and queuing models. Owing to the interaction of two countervailing effects, utility maximization does not lead to corner solutions. For example, an inventor can decrease excess examination time by choosing a low-workload filing strategy. However, this strategy also increases deviation costs. Additionally, if many other inventors also choose low-workload filing strategies, their patent applications will also reach standard examination time early, which increases excess examination time. Thus, whereas inventors can directly influence deviation costs, costs of excess examination time depend on the choices of all inventors in the cohort and are affected only indirectly

4 Dynamic patent congestion

To model congestion at the patent office we adapt the dynamic bottleneck model introduced by Vickrey (1969) and generalized by Small and Verhoef (2007). Consider a cohort of patent applications filed at date $t_f = 0$. Desired pendency durations t_p lie between a lower bound $\underline{t}_p$ and an upper bound $\overline{t}_p$ with $\underline{t}_p \geq \underline{t} \geq t_f$ where $\underline{t}$ is the date after which patent applications in the cohort face excess examination time. Let $\overline{N}$ denote the total number of patent applications in the cohort; $N(t)$, the number of patent applications that have not reached standard examination time until t; and $N_e(t)$ the number of examined applications at time t. The number of patent applications in the backlog is thus $\overline{N} - N_e(t)$. Although all applications wait for examination from their filing date onward, they formally "reach the bottleneck" only after their standard examination time has elapsed. The flow of patent applications reaching their standard examination time, $s = \partial N(t)/\partial t$, will be determined endogenously by the filing strategies of inventors. Parameter s is thus an average flow that depends on inventors' heterogeneity with respect to desired pendency duration. The specific composition of this parameter will be derived in Section 4.2. All patent applications in the cohort have reached their standard examination time when $N(t) = 0$. Thus it must hold that

$$\overline{N} = \int_{\underline{t}_p}^{\overline{t}_p} s \, dt = s(\overline{t}_p - \underline{t}_p). \tag{1}$$

The examination capacity of the patent office is κ examination steps per time interval. All applications have to pass this bottleneck. Define $B(t)$ as the number of patent applications "in the queue," that is, applications delayed by the bottleneck facing excess examination time in t. The relation of applications reaching standard examination time and prosecuted applications is given by the kinked function

$$n_e = \begin{cases} s & \text{if } \; s \leq \kappa \; \text{ and } \; B(t) = 0 \; \text{ for } \; \underline{t}_p \leq t \leq \overline{t}_p \\ \kappa & \text{otherwise.} \end{cases}$$

where n_e defines the flow of patent applications being issued: $\partial N_e(t)/\partial t = n_e$. Assumption 4 that the patent office faces a backlog means that $n_e = \kappa$. Focusing on the current cohort only and ignoring any preexisting backlog we have

$$B(0) = 0.$$

by a single inventor's filing strategy. Another advantage over queuing models is that the arrival rate at the bottleneck is endogenously determined by inventors who strategically choose their filing strategies. In contrast, the arrival rate in queuing models is given exogenously.

Then

$$\frac{\partial B(t)}{\partial t} = s - \kappa \ \forall \ t > 0$$

is the marginal change of patent applications facing excess examination time. Time $\overline{t}$ when the backlog caused by the cohort vanishes is defined by

$$B(\overline{t}) = \int_{\underline{t}}^{\overline{t}} (s - \kappa) dt = 0. \tag{2}$$

At any time $\underline{t} \leq t \leq \overline{t}$ the number of applications that have reached their standard examination time is given by

$$B(t) = \int_{\underline{t}}^{t} (s - \kappa) dz. \tag{3}$$

Given the bottleneck's capacity, at some time $t > \underline{t}$ the backlog causes excess examination time of $B(t)/\kappa = \int_{\underline{t}}^{t} (s/\kappa - 1) dz$. The excess examination time $X(t)$ of a patent application that reaches its standard examination time in t thus amounts to

$$X(t) = \left[\frac{s}{\kappa} - 1\right](t - \underline{t}). \tag{4}$$

Figure 1 provides the intuition behind equation (4). In Figure 1(a) the flow of patent applications reaching their standard examination time is $s = 2$ applications per time interval whereas the capacity of the bottleneck is $\kappa = 1$. Using equation (4) we can calculate the total excess examination time an application faces. For instance, given that P5 has a standard examination time of 3 time intervals, P5's examination duration is prolonged by $X(3) = \left[\frac{2}{1} - 1\right] 3 = 3$ intervals of excess examination time. In Figure 1(b) the flow of applications reaching their standard examination time increases to $s = 3$. Hence application P5's excess examination time now increases to $X(3) = \left[\frac{3}{1} - 1\right] 3 = 6$ time intervals.

4.1 Quantifying the costs of patent pendency

Define the filing strategy of a patent application as δ, meaning that this patent application has a standard examination time of $t(\delta)$ where $\partial t(\delta)/\partial \delta > 0$. It follows that the choice of δ is equivalent to the choice of $t(\delta)$. For ease of exposition, we use $t(\delta) \equiv t_\delta$ in the following.

Costs of excess examination time

By choosing their filing strategy, δ, inventors can directly and indirectly influence the excess examination time of their patent application. The direct path is the choice of the application's arrival in the queue of applications with excess examination time. The indirect path is one application's effect on overall excess examination time. All $\overline{N}$ inventors face excess examination

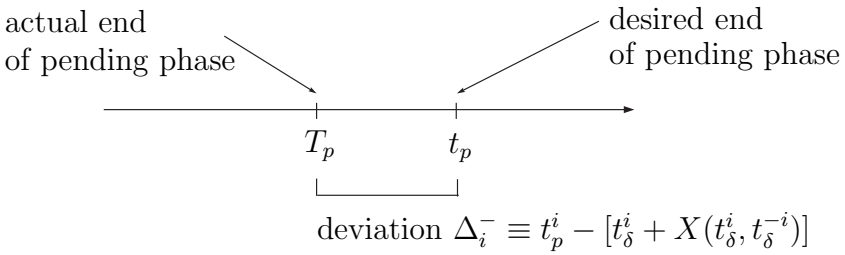


Figure 2: Hastened prosecution

costs of $\alpha X(t_\delta^i, t_\delta^{-i})$, where superscript $-i$ indicates all other inventors in the cohort and α quantifies the magnitude of these costs.

Costs of deviating from desired pendency

In addition to excess examination costs, inventors also face deviation costs that depend on the deviation between the desired and actual length of pendency duration. Actual pendency duration T_p^i is given by standard examination time t_δ^i, plus the excess examination time caused by the backlog, $X(t_\delta^i, t_\delta^{-i})$, that is,

$$T_p^i = t_\delta^i + X(t_\delta^i, t_\delta^{-i}). \tag{5}$$

To reduce excess examination time inventors can adjust their filing strategies by either reducing or increasing the workload associated with their patent applications. While choosing a filing strategy incorporating fewer (or more) examination steps comes at the cost of deviating from desired pendency, it may lead to a higher utility because excess examination time is reduced. Typically, inventors can instruct their patent attorneys to adapt the drafting of the patent document in such a way as to require fewer (or more) examination steps and thus shorten (or prolong) the standard examination time of their patent application.

Figure 2 depicts the computation of shortening pendency resulting in $T_p^i < t_p^i$. The inventor faces deviation costs contingent on the extent of the deviation $\nu^h \Delta_i^-$, where $\Delta_i^- \equiv t_p^i - [t_\delta^i + X(t_\delta^i, t_\delta^{-i})]$ and ν^h quantifies the magnitude of these costs.

Figure 3 depicts the computation of delaying prosecution (prolonging pendency) resulting in $T_p^i > t_p^i$. The inventor faces deviation costs contingent on the extent of the deviation, $\nu^d \Delta_i^+$, where $\Delta_i^+ \equiv t_\delta^i + X(t_\delta^i, t_\delta^{-i}) - t_p^i$ and ν^d quantifies the magnitude of these costs.

Inventors in the cohort are heterogeneous regarding the desired duration of pendency. Simultaneous utility maximization will lead to one group hastening the prosecution process and

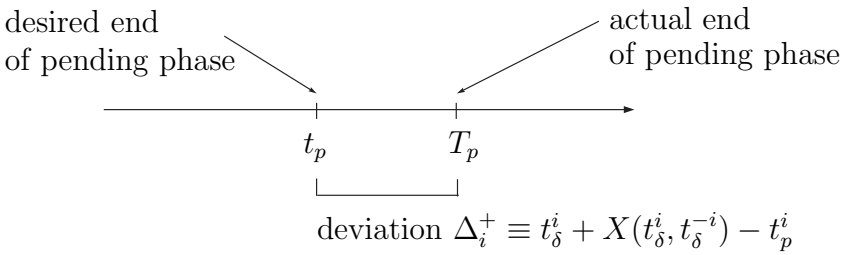


Figure 3: Delayed prosecution

the other delaying it (see below). Inventors in the first group thus face deviation costs $\nu^h \Delta_i^-$, while inventors in the latter group face deviation costs $\nu^d \Delta_i^+$.

Trade-off between excess examination time and deviation

In choosing their optimal filing strategy inventors face a tradeoff between deviation costs and excess examination time. To avoid excess examination time inventors may choose low(high)-workload strategies, however, such strategic manipulation causes deviation costs that could outweigh the gains to be derived by reducing excess examination time. Additionally, while the choice of a better filing strategy for a single inventor could lower her excess examination time, the effect diminishes as more inventors adopt the same strategy. Indeed, if many inventors choose a better filing strategy, their lower standard examination times increase the flow of patent applications reaching excess examination time (s). This, in turn, increases overall excess examination time, $\partial X/\partial s > 0$. Thus, inventors need to balance these opposing forces when choosing their optimal filing strategy.

Summarizing, we can specify the utility function of an inventor i as

$$U_i(t_\delta^i, t_\delta^{-i}) = \begin{cases} \bar{u} - \alpha X(t_\delta^i, t_\delta^{-i}) - \nu^h \, \Delta_i^-(t_\delta^i, t_\delta^{-i}) - \varphi & \text{if } T_p^i < t_p^i \\ \bar{u} - \alpha X(t_\delta^i, t_\delta^{-i}) - \varphi & \text{if } T_p^i = t_p^i \\ \bar{u} - \alpha X(t_\delta^i, t_\delta^{-i}) - \nu^d \, \Delta_i^+(t_\delta^i, t_\delta^{-i}) - \varphi & \text{if } T_p^i > t_p^i, \end{cases} \tag{6}$$

where variable $\bar{u}$ captures the utility from patenting (which by assumption exceeds costs to reflect the nonelastic demand for patenting) and φ reflects the administrative fees.

Equilibrium occurs when no inventor can increase her utility by deviating from her optimal filing strategy. Inventors achieve an early decision date, $T_p^i < t_p^i$, by choosing a low-workload filing strategy and a late decision date, $T_p^i > t_p^i$, by choosing a high-workload filing strategy.

4.2 Filing strategies in equilibrium

This section derives the aggregate dynamic equilibrium for which no single inventor has an incentive to change her filing strategy. The aim is to identify the parameters driving the cutoff threshold dividing inventors into the groups of (i) inventors choosing low-workload filing strategies; and (ii) inventors choosing high-workload filing strategies.

No deviation from desired pendency duration

An inventor who has no deviation from her desired pendency duration, $T_p^i = t_p^i$, has the utility function $\bar{u} - \alpha X(t_\delta^i, t_\delta^{-i}) - \varphi$. Maximization with respect to filing strategy t_δ^i yields $\partial X(t_\delta^i, t_\delta^{-i})/\partial t_\delta^i = 0$ meaning that utility is maximal when her patent application's standard examination time causes no change in excess examination time.

Delayed prosecution (prolonged pendency)

All inventors with $T_p^i > t_p^i$ face utility $\bar{u} - \alpha X(t_\delta^i, t_\delta^{-i}) - \nu^d \left[t_\delta^i + X(t_\delta^i, t_\delta^{-i}) - t_p^i\right] - \varphi$ (from equation 6). Maximizing utility with respect to the chosen filing strategy t_δ^i yields the first-order condition

$$\frac{\partial X(t_\delta^i, t_\delta^{-i})}{\partial t_\delta^i} = \frac{-\nu^d}{\alpha + \nu^d}, \quad \forall\, i \quad \text{if} \quad T_p^i > t_p^i \tag{7}$$

which gives an implicit solution for the individually optimal filing strategy t_δ^i in case of $T_p^i > t_p^i$. As long as this condition holds, an inventor delaying prosecution (i.e., prolonging pendency) maximizes the utility to be derived from patenting.

Hastened prosecution (shortened pendency)

All types with $T_p^i < t_p^i$ face utility $\bar{u} - \alpha X(t_\delta^i, t_\delta^{-i}) - \nu^h \left[t_p^i - t_\delta^i - X(t_\delta^i, t_\delta^{-i})\right] - \varphi$. Again, inventors maximize utility with respect to their filing strategy t_δ^i and the first-order condition is

$$\frac{\partial X(t_\delta^i, t_\delta^{-i})}{\partial t_\delta^i} = \frac{\nu^h}{\alpha - \nu^h} \quad \forall\, i \quad \text{if} \quad T_p^i < t_p^i \tag{8}$$

which gives an implicit solution for the individually optimal filing strategy t_δ^i in case of $T_p^i < t_p^i$. As long as this condition holds, an inventor shortening pendency maximizes the utility to be derived from patenting.

Excess examination time

The resulting marginal change in excess examination time with respect to chosen standard examination time, $\partial X(t_\delta^i, t_\delta^{-i}))/\partial t_\delta^i$, is positive for inventors shortening pendency (equation 8)

and negative for inventors prolonging pendency (equation 7). For hastened pendency this observation means that applications presenting a higher workload face higher excess examination time, whereas for prolonged pendency it means that applications with a higher workload face lower excess examination time. Thus starting in $\underline{t}$ after the filing date of the applications in the cohort, excess examination time first increases and later, as more and more applications with hastened pendency are prosecuted, decreases again.

To derive equilibrium strategies one must take into account the patent office's restricted capacity. From equation (4), the marginal change of excess examination time resulting from the restricted examination capacity of the office can be calculated as

$$\frac{\partial X(t)}{\partial t} = \frac{s}{\kappa} - 1. \tag{9}$$

Combining this result with the implicit solution for optimal filing strategies in case of hastened pendency (equation 8) yields the flow of low-workload patent applications reaching excess examination time

$$s^{\text{low}} = \frac{\alpha}{\alpha - \nu^h}\kappa. \tag{10}$$

This flow guarantees maximal utility for all inventors shortening pendency, meaning that they have no incentive to change their filing strategies.

Analogously, the flow of patent applications by inventors prolonging pendency (equation 7) reaching standard examination time is given by

$$s^{\text{high}} = \frac{\alpha}{\alpha + \nu^d}\kappa. \tag{11}$$

This flow guarantees maximal utility for inventors prolonging pendency, meaning that they have no incentive to change their filing strategies.

Between these two groups lies the "cutoff" inventor, who does not deviate from desired pendency duration because her actual and desired pendency duration are the same, $\hat{T}_p = \hat{t}_p$. She stands on the line that divides the cohort into two parts: (i) inventors hastening the prosecution process and (ii) inventors prolonging it. We denote her filing strategy as $\hat{t}_\delta$, and use $\hat{X}_\delta$ to refer to the maximal excess examination time that she faces.[5] Using equation (5) we can specify her filing strategy as

$$\hat{t}_\delta = \hat{T}_p - \hat{X}_\delta. \tag{12}$$

Note that the flow of applications reaching their standard examination time is higher for those with low-workload filing strategies than for those with high-workload strategies—that

[5]Recall that maximization of the utility for an inventor with $T_p^i = t_p^i$ yields $\partial X/\partial t = 0$.

is, $s^{\text{low}} > s^{\text{high}}$. For part (i) of the cohort, therefore, the flow of patent applications reaching their standard examination time is higher than for part (ii). Given the above results, the flow of low-workload applications reaching standard examination time leads to an increasing number of patent applications facing excess examination time before maximal excess examination time is incurred by the cutoff inventor. The lower flow of high-workload filings reaching standard examination time then leads to a decreasing number of patent applications facing excess examination time after maximal excess examination time is incurred by the cutoff inventor. Consequently all inventors in part (i) of the cohort choose lower-workload filing strategies than the cutoff inventor, whereas all inventors in part (ii) of the cohort choose higher-workload filing strategies than the cutoff inventor.[6]

During the interval $[\underline{t}, \overline{t}]$, that is, from the date on which patent applications face excess examination time until the last patent application is issued, the patent office issues patents according to its capacity κ. Thus, using equation (1) it must be that

$$\overline{t} - \underline{t} = \overline{N}/\kappa = s(\overline{t}_p - \underline{t}_p)/\kappa. \tag{13}$$

Consequently, as long as $s > \kappa$ the backlog exists even past the longest preferred pendency duration in the distribution, $\overline{t}_p$. We derive below the composition of the cohort with respect to both types.

Share of low-workload applications

As a benchmark we calculate the share of low-workload applications. Recall that a patent application faces excess examination time after its standard examination time has elapsed. If many patent applications are written in a clear and concise manner, many will reach this point early. If many applications are imprecise, they cause a higher workload but face excess examination time later because they have a longer standard examination time. In a first step we calculate the proportion of patent applications facing excess examination time up to the point in time when the desired pendency of the cutoff inventor, $\hat{t}_p$, is reached. Until $\hat{t}_p$ a number of $\sigma\overline{N} = \int_{\underline{t}_p}^{\hat{t}_p} s\,dt$ patents faces excess examination time, where $0 \leq \sigma \leq 1$. Inserting $\overline{N}$ from equation (1) yields the proportion of patents facing excess examination time before the cutoff

[6]Due to the linear utility function inventors in part (i) {part (ii)} are in fact indifferent among filing strategies with standard examination times $[\underline{t}, \hat{t}_\delta]$ $\{[\hat{t}_\delta, \overline{t}]\}$. However, following Small and Verhoef (2007), we assume—without loss of generality—that patent applications, once they face excess examination time and thus arrive in the queue, are examined in the order of their desired pendency duration dates.

inventor,

$$\sigma = \frac{\hat{t}_p - \underline{t}_p}{\overline{t}_p - \underline{t}_p}. \tag{14}$$

This proportion gives us the size of part (i) of the cohort subject to desired pendencies $\underline{t}_p$, $\overline{t}_p$, and $\hat{t}_p$. However, the value of $\hat{t}_p$, the desired pendency marking the cutoff inventor, is subject to inventors' filing strategies. In part (i) of the cohort the flow of inventors facing excess examination time is s^{low} (equation 10). The first patent application faces excess examination time at $\underline{t}$, whereas the cutoff inventor faces excess examination time at $\hat{t}_\delta$, when her application's standard examination time elapses. Given this, the benchmark share of low-workload applications can be calculated as

$$\sigma^{low}\overline{N} = \int_{\underline{t}}^{\hat{t}_\delta} s^{\text{low}} dt = \frac{\alpha\kappa}{\alpha - \nu^h}(\hat{t}_\delta - \underline{t}). \tag{15}$$

Derivation of equilibrium values

Equilibrium occurs when inventors have no incentives to manipulate their filing strategies in order to shorten or prolong pendency duration. We identify it using the filing strategy of the cutoff inventor, who should be indifferent between choosing her optimal filing strategy with no deviation and deviating by either choosing a lower or higher workload filing strategy.

We show in Appendix A that given the cutoff inventor's indifference, all inventors facing shorter (longer) pendency duration have no incentive to deviate either. In such an equilibrium, σ^{low} is the proportion of inventors choosing a lower-workload filing strategy than the cutoff inventor and can be calculated as

$$\sigma^{low} = \frac{\nu^d}{\nu^d + \nu^h}. \tag{16}$$

In equilibrium the cutoff inventor faces excess examination time

$$\hat{X}_\delta = \frac{\nu^d \nu^h \overline{N}}{(\nu^d + \nu^h)\alpha\kappa}. \tag{17}$$

The first patent application faces excess examination time in

$$\underline{t} = \underline{t}_p - \sigma^{low}(\overline{t}_p - \underline{t}_p)\left[\frac{s}{\kappa} - 1\right], \tag{18}$$

whereas the date at which the backlog vanishes is given by

$$\overline{t} = \overline{t}_p + (1 - \sigma^{low})(\overline{t}_p - \underline{t}_p)\left[\frac{s}{\kappa} - 1\right]. \tag{19}$$

Note that the proportion σ^{low} of inventors choosing a low-workload filing strategy does not depend on the costs of excess examination time α, but on the parameters ν^d and ν^h, which drive deviation costs:

- *The influence of hastening costs ν^h.* If deviation costs for low-workload inventors increase, the proportion of low-workload applications decreases. To maintain indifference as hastening becomes relatively less attractive the cutoff inventor's "location" moves to the left, $\partial \hat{t}_p / \partial \nu^h < 0$, which narrows the range for low-workload applications. The proportion σ^{low} thus decreases as fewer patents face excess examination time before the cutoff inventor does.

- *The influence of delaying costs ν^d.* If deviation costs for high-workload inventors increase, the proportion of low-workload applications increases. To maintain indifference as delaying becomes relatively less attractive the cutoff inventor's "location" moves to the right, $\partial \hat{t}_p / \partial \nu^d > 0$. Therefore the range of low-workload applications increases—that is, more patent applications face excess examination time before the cutoff inventor does.

5 Policy responses to the backlog

This section puts the model to work to study possible policy responses to the backlog. It presents the welfare function, provides an estimate for σ^{low}, and discusses three policy responses.

5.1 Welfare function

The welfare function considers the impact on society (S), the patent office (O), and inventors (I). We write it as

$$\Omega = \underbrace{-(1-\sigma^{low})\Phi}_{S} \underbrace{-w(\kappa)(\overline{t}-\overline{t}_p)}_{O} \underbrace{-\sigma^{low}(\beta+\gamma^h)-(1-\sigma^{low})(\beta+\gamma^d)}_{I}, \tag{20}$$

where β reflects the average costs of excess examination time and γ^d (γ^h) the average deviation costs caused by longer- (shorter-) than-desired pendency duration. We explain the three elements of the welfare function in turn.

Impact on society

Component (S) captures social damages caused by applications with high-workload filing strategies. The parameter Φ is a theoretical construct that captures the social costs if *all* patents had high-workload filing strategies. We refer to this parameter as the "costs of hypercongestion."

As explained in Section 2, patent pendency is associated with uncertainty about the status of property rights, which may defer the introduction of new products to the market and distort

rival firms' investment decisions. We are not aware of peer-reviewed estimates of Φ. One study, commissioned by the United Kingdom's Intellectual Property Office, estimates that a one-year increase in pendency at the trilateral offices is associated with a cost of £7.6 billion per annum on the global economy (London Economics, 2010, p. viii)—an amount equivalent to the yearly R&D expenditures of Switzerland. Although this estimate is subject to much uncertainty, our welfare analysis adopts the dominant view that the backlog has overall negative social effects.

Given the proportion of low-workload applications, σ^{low}, actual damages amount to $(1 - \sigma^{low})\Phi$. If Φ is high, policy measures aimed at increasing the proportion of low-workload applications will have a particularly strong impact on welfare. Table 1 summarizes the comparative statics concerning σ^{low}. Naturally, an increase in the proportion of low-workload applications has a positive welfare effect: $\partial S/\partial \sigma^{low} > 0$.

Impact on the patent office

The second component (O) captures the costs borne by the patent office for the additional examination time caused by the backlog. Recall that the backlog vanishes after the latest desired pendency duration is reached, $\overline{t} > \overline{t}_p$. The cost induced by the backlog for the patent office is the wage $w(\kappa)$ of patent examiners multiplied by additional examination time $\overline{t} - \overline{t}_p$. From the revenue side, the patent office receives constant filing fees φ per patent application in the cohort. Fees are transfer payments from inventors to the patent office and, therefore, they do not appear in the welfare function. Note that the backlog does not affect filing fees. However, we will discuss backlog-specific fees in Response 2 below.

The additional examination time caused by the backlog is of central interest to the patent office. In equilibrium the date on which the backlog vanishes, $\overline{t}$, is given by equation (19) and depends on deviation costs. From this, it is straightforward to calculate additional examination time $(\overline{t} - \overline{t}_p)$

$$\overline{t} - \overline{t}_p = (1 - \sigma^{low})(\overline{t}_p - \underline{t}_p) \left[\frac{s}{\kappa} - 1\right]. \tag{21}$$

We discussed earlier that the proportion of low-workload applications σ^{low} increases with the cost of prolonging pendency ν^d. Therefore, an increase in ν^d decreases additional examination time by reducing the office's overall workload. The situation regarding the costs of shortening pendency ν^h is reversed. An increase of ν^h leads to an increase in additional examination time since a greater number of high-workload applications increases overall workload, $\partial(\overline{t} - \overline{t}_p)/\partial \nu^h > 0$. Table 1 summarizes the comparative statics regarding $\overline{t} - \overline{t}_p$.

Impact on inventors

Component (I) includes the costs and benefits of excess examination time as well as deviation costs faced by inventors. All inventors face costs $\beta(t_\delta^i, t_\delta^{-i})$ of excess examination time. In addition, inventors with a low-workload filing strategy face deviation costs $\gamma^h(t_\delta^i, t_\delta^{-i})$ whereas inventors with a high-workload filing strategy face deviation costs $\gamma^d(t_\delta^i, t_\delta^{-i})$.

We calculate the average costs of excess examination and of deviation from desired pendency following Small and Verhoef (2007). Given the uniform distribution of desired pendency durations, costs of excess examination time per inventor increase monotonically from zero to their maximum at $\hat{t}_p$ and then decrease monotonically to zero. Therefore we compute average excess examination costs per inventor by dividing the sum of the minimum (which is zero and is realized by the inventors who present the lowest and highest workload) and maximum value by two, which yields $\beta = \alpha\hat{X}_\delta/2$. Inserting $\hat{X}_\delta$ from equation (17) we have

$$\beta = \frac{\nu^h \sigma^{low} \overline{N}}{2\kappa}. \tag{22}$$

The average costs of excess examination decrease with the capacity of the patent office and increase with hastening costs, the proportion of low-workload applications, and the number of patent applications in the cohort.

Average deviation costs can be calculated analogously: deviation costs are given by $\nu^\tau \Delta$, $\tau = d, h$, where Δ is defined as the deviation between actual and desired pendency duration. They reach their maximum at Δ^-_{max} and Δ^+_{max}, respectively. Because $\Delta^- \equiv t_p - T_p$ we have Δ^-_{max} for the inventor with the lowest-workload filing with examination date $T_p = \underline{t}$. This inventor faces maximal deviation costs, $\nu^h(\underline{t}_p - \underline{t})$, but no excess examination costs. For later examination dates, deviation costs decrease to zero for the cutoff inventor, and, from there, they increase back to a maximum at Δ^+_{max} for the inventor with the highest-workload filing with examination date $T_p = \overline{t}$. This inventor faces maximal deviation costs $\nu^d(\overline{t} - \overline{t}_p)$, but no excess examination costs. Both the lowest- and the highest-workload inventors would be worse off if they chose a higher (lower) workload filing strategy, which they would do by trading less deviation costs for higher excess examination costs.

Using equation (19) we calculate the average deviation costs per inventor for low-workload applications by dividing the sum of the minimum (which is zero and is realized by the cutoff inventor) and maximum value by two, which yields

$$\gamma^h = \frac{\nu^h \sigma^{low} \overline{N}}{2\kappa} \left[1 - \frac{\kappa}{s}\right]. \tag{23}$$

Using equation (18) we calculate average deviation costs per inventor for high-workload

Table 1: Comparative statics of the proportion of low-workload applications, additional examination time, and inventors' average costs

	σ^{low}	$\overline{t} - \overline{t}_p$	β	γ^d	γ^h
ν^d	+	–	+	+	+
ν^h	–	+	+	+	+

applications by dividing the sum of the minimum (which is zero and is realized by the cutoff inventor) and maximum value by two, which yields

$$\gamma^d = \frac{\nu^d(1-\sigma^{low})\overline{N}}{2\kappa}\left[1-\frac{\kappa}{s}\right]. \tag{24}$$

Average deviation costs obviously increase in ν^h and ν^d. Table 1 summarizes comparative statics. Straightforward calculations show that $\gamma^d = \gamma^h$. Given this, we can rewrite the costs to inventors as $I \equiv -(\beta + \gamma)$.

5.2 Estimate of σ^{low}

The parameter σ^{low} is an important determinant of the welfare function. As far as we can ascertain no data exist that would allow us to compute σ^{low}. Therefore, we obtained an estimate of σ^{low} from seven current and former chief economists and economic advisors at six patent offices (CIPO, EPO, IP Australia, JPO, USPTO, WIPO). Chief economists are particularly well informed about σ^{low}: they frequently interact with patent applicants, they are data proficient, and they have a fair understanding of the procedural techniques that patent attorneys use to affect pendency duration.[7]

For modeling purposes the theoretical model distinguishes between low-workload patent applications seeking to hasten the grant decision and high-workload patent applications seeking to postpone it. It neglects the fact that some applicants do nothing to affect the speed of issuance. Therefore we also asked chief economists for the proportion of applications for which inventors take no action to affect pendency duration. The exact questions appear in Appendix B. Figure 4 reports the anonymized answers. According to the experts, the proportion of

[7]Two alternative approaches for estimating σ^{low} involve asking patent applicants directly and using administrative data to identify actions taken to accelerate or delay the examination process. The first approach is expensive to implement and relies on applicants telling the truth. The second approach would be more factual but it would provide us with office-specific "endogenous" rates, that is, rates conditional on the existing pendency duration and the existing structure of incentives at patent offices. Our approach of aggregating expert opinions lies between these two extremes.

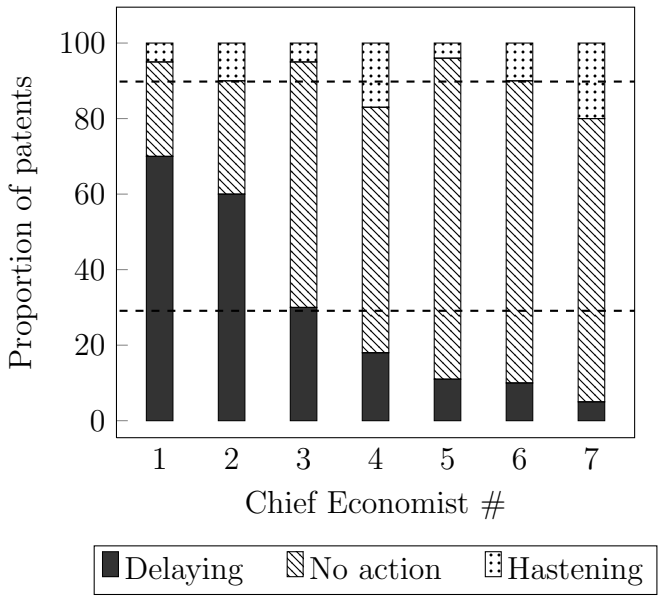


Figure 4: Seven chief economists' assessments of σ^{low}

applications for which inventors seek to increase pendency duration, $(1 - \sigma^{low})$, averages 0.29 (ranging from a minimum of 0.05 to a maximum of 0.70). The proportion of applications for which inventors seek to expedite the process, σ^{low}, is much lower, with a mean score of 0.10 (from a minimum of 0.05 to a maximum of 0.20). Only one expert believes that the proportion of applications for which inventors seek to expedite the process exceeds the proportion of applications for which inventors seek to delay it.

Distributing evenly between the two groups those applications for which no intentional action is taken to hasten or delay (the remaining 61 percent) gives an overall σ^{low} share of 0.405 (=0.10 + 0.61/2). Distributing these applications on a pro-rata basis gives an overall σ^{low} share of 0.258 (=0.10+0.61*0.10/0.39). It goes without saying that estimates of σ^{low} are conditional on the specific situation and incentives at the corresponding patent office. The actual mean of σ^{low} across offices is therefore of little practical use—the only insight that we rely on in the development of policy responses is that experts believe that intentional delays exceed intentional accelerations.[8] We now turn to discussing three policy responses to the backlog.

[8]Hard data obtained from the Swiss IP Institute reinforce this conclusion. Applicants have the option to ask for accelerated examination or to postpone it. Respectively 6 and 14 percent of applicants made such request.

5.3 Response 1: Increase the patent office's examination capacity

Increasing the examination capacity reduces excess examination and deviation costs for inventors but has an ambiguous effect for the patent office. The office faces an increase in wages but a decrease in additional examination time caused by the backlog (that is, $\overline{t} - \overline{t}_p$ decreases). The overall effect can be calculated as

$$\frac{\partial \Omega}{\partial \kappa} = \underbrace{-\frac{\partial w}{\partial \kappa}(\overline{t} - \overline{t}_p) - \frac{\partial (\overline{t} - \overline{t}_p)}{\partial \kappa} w(\kappa)}_{dO} \underbrace{- \frac{\partial \beta}{\partial \kappa} - \frac{\partial \gamma}{\partial \kappa}}_{dI} \tag{25}$$

where part dO captures the countervailing effects at the patent office and part dI captures the effects for inventors. Whereas part dI has an unambiguously positive impact on welfare because $\frac{\partial \beta}{\partial \kappa} < 0$ and $\frac{\partial \gamma}{\partial \kappa} < 0$, the sign of part dO is ambiguous.

Increased capacity could decrease the office's expenses owing to less additional examination time, $\partial(\overline{t} - \overline{t}_p)/\partial \kappa < 0$. However, increasing processing capacity is costly, $\frac{\partial w}{\partial \kappa}(\overline{t} - \overline{t}_p) > 0$. The office must either force existing examiners to work overtime or hire additional staff. Part dO in equation (25) is positive if and only if the cost of increasing processing capacity is outweighed by shorter additional examination time.

Inserting additional examination time from equation (21), the welfare effect for the patent office can be specified as

$$dO = (1 - \sigma^{low})(\overline{t}_p - \underline{t}_p) \left[-\frac{\partial w}{\partial \kappa} \frac{(s - \kappa)}{\kappa} + \frac{s\, w(\kappa)}{\kappa^2} \right]. \tag{26}$$

This effect is positive whenever

$$\frac{s}{s - \kappa} > \varepsilon_{w,\kappa} \tag{27}$$

where $\varepsilon_{w,\kappa} \equiv (\partial w/\partial \kappa)/(w(\kappa)/\kappa)$ is the capacity elasticity of wages. If an increase of 1 percent in the pool of examiners leads to a concomitant 1-percent increase in both examination capacity and the wage bill, we would have $\varepsilon_{w,\kappa} = 1$. In addition, as long as the office is congested we have $s/(s - \kappa) > 1$. Thus, increasing capacity in case of $\varepsilon_{w,\kappa} \leq 1$ has a positive effect.

However, the effectiveness of this policy response—that is, the extent of its positive impact on welfare—depends on σ^{low}. Differentiating dO with respect to the proportion of low-workload applications yields

$$\frac{\partial dO}{\partial \sigma^{low}} = \frac{\partial w}{\partial \kappa}(\overline{t}_p - \underline{t}_p) \left[\frac{s}{\kappa} - 1 \right] - (\overline{t}_p - \underline{t}_p) \frac{s}{\kappa^2} w(\kappa). \tag{28}$$

It is straightforward to show that a greater proportion of low-workload applications has a positive effect, $\partial dO/\partial \sigma^{low} > 0$, whenever condition (27) is fulfilled. Thus, given a low wage

elasticity, the positive welfare effect of a higher examination capacity increases with the proportion of low-workload applications. This leaves us to investigate the effect of σ^{low} on the extent of the positive welfare effect for inventors, dI. Inserting the derivatives of deviation costs (equations 23 and 24) we can specify the welfare effect as $dI = [\nu^d + \sigma^{low}(\nu^h - \nu^d)]\overline{N}/\kappa^2$. This quantity increases in σ^{low} whenever shortening pendency is more costly than prolonging pendency, $\nu^h > \nu^d$. Putting these results together we have:

Result 1 *Increasing the patent office's capacity increases welfare if and only if the wage elasticity regarding the capacity increase is sufficiently low. Where that condition is present, an increase in low-workload applications increases the effectiveness of this policy response whenever shortening pendency is more costly than prolonging pendency.*

The cost of—and opportunity for—increasing the labor force varies by office. Picard and van Pottelsberghe de la Potterie (2013) report that the workload of examiners at the JPO and the EPO is less than 50 percent of the workload of US examiners. They also report evidence that the JPO and the EPO offer more attractive compensations to their examiners than does the USPTO. Thus, it would seem that there is more scope for increasing capacity at the USPTO than at the JPO and the EPO. However, expert estimates summarized in Figure 4 suggest that the proportion of low-workload applications is rather low, suggesting that this policy response may not be very effective. While a sufficiently low wage elasticity would yield a positive effect on office expenses (part dO), the positive effect on inventors would be small as long as the costs of prolonging pendency exceeded those of shortening pendency. Nonetheless, given a low wage elasticity increasing examination capacity has a positive welfare effect.

5.4 Response 2: Introduce penalty fees

The findings regarding the interplay of deviation costs (ν^d, ν^h) and average excess examination costs caused by the backlog (β) suggest that one can manipulate deviation costs to increase welfare. One way to achieve this result is to introduce penalty fees.[9] Penalty fees may take the form of claim-based fees or page-based fees, which are already used by many patent offices.

There are two ways to structure penalty fees. They can increase linearly with the number of claims or pages, a method that we call the *linear penalty fee*. Alternatively, they can target high-workload applications only by identifying thresholds at which penalties are incurred (the

[9] As the introduction of penalty fees is a pure transfer between inventors and the office, the payments themselves do not appear in the welfare function.

stepped-penalty fee). Although a linear penalty fee affects both, low- and high-workload applications, it targets features of the patent document or the examination process that impose extra workload and thus predominantly affect high-workload applications. A priori, there are good intuitive reasons for both types of fees. Analytically, both types influence the cost of deviating from desired pendency duration, but to a different extent.

Calculating the derivative of the welfare function (20) with respect to ν^τ, where τ is a placeholder equal to d or h, yields

$$\frac{\partial \Omega}{\partial \nu^\tau} = \frac{\partial \sigma^{low}}{\partial \nu^\tau}\Phi - w(\kappa)\frac{\partial(\overline{t} - \overline{t}_p)}{\partial \nu^\tau} - \frac{\partial \beta}{\partial \nu^\tau} - \frac{\partial \gamma}{\partial \nu^\tau}. \tag{29}$$

The last two terms impose negative effects, but the first two terms show differing signs. For hastening costs, $\partial \sigma^{low}/\partial \nu^h$ is negative, but $\partial(\overline{t} - \overline{t}_p)/\partial \nu^h$ is positive. For delaying costs, $\partial \sigma^{low}/\partial \nu^d$ is positive, but $\partial(\overline{t} - \overline{t}_p)/\partial \nu^d$ is negative. With these elements in mind, we can study both types of fees.

Stepped penalty fee

A stepped penalty fee increases the costs of prolonging pendency (ν^d) without affecting the costs of shortening pendency (ν^h). The welfare effect is positive if the cost of hypercongestion is sufficiently high. Indeed, imposing $\partial \Omega/\partial \nu^d > 0$ yields a critical threshold for the potential damage Φ:

$$\Phi > \left[\frac{\partial(\overline{t} - \overline{t}_p)}{\partial \nu^d} w(\kappa) + \frac{\partial \beta}{\partial \nu^d} + \frac{\partial \delta}{\partial \nu^d}\right] \left(\frac{\partial \sigma^{low}}{\partial \nu^d}\right)^{-1} \equiv \Phi_d.$$

Thus, a stepped penalty fee has a positive welfare effect whenever the costs of hypercongestion are higher than the threshold value Φ_d.

Linear penalty fee

A linear penalty fee, ξT_p, $\xi > 0 \; \forall \; t_\delta$, increases the costs of prolonging pendency. The effect of a linear penalty fee on the costs of shortening pendency depends on the size of the penalties. If a low-workload inventor chooses fewer examination steps (that is, higher deviation Δ^-) this affects the penalty fee ξT_p that she has to pay, but also her deviation costs $v^h \Delta^-$: While deviation costs increase as the standard examination time decreases, $\partial \nu^h \Delta^-/\partial t_\delta < 0$, the penalty fee decreases because it punishes higher workloads and "rewards" lower ones, $\partial \xi T_p/\partial t_\delta > 0$. In order to compute the overall effect we need to specify which effect is stronger. For $\left|\frac{\partial \xi T_p}{\partial t_\delta}\right| > (<) \left|\frac{\partial \nu^h \Delta^-}{\partial t_\delta}\right|$ the penalty-fee reduction achieved by choosing a lower-workload filing strategy is more (less) prominent than the resulting increase in deviation costs. This relation obviously depends on

the relative magnitude of the penalty fee versus deviation costs, $\xi > (<)\nu^h$. We therefore need to distinguish both possible cases—moderate fees ($\xi < \nu^h$) and high fees ($\xi > \nu^h$)—when discussing the welfare effect of a linear penalty fee.

Let us first consider the case of moderate fees, $\xi < \nu^h$. In this case, lowering the standard examination time has a negative effect overall, so the introduction of a linear penalty fee has the same effect as an increase in deviation costs ν^h. Given the welfare effect of changing ν^h discussed above (equation 29), increasing the costs of hastening prosecution decreases welfare. In case of low costs of hypercongestion ($\Phi < \Phi_d$), the overall welfare-effect of a linear penalty fee would thus be negative.

Let us now turn to the case of high fees, $\xi > \nu^h$, where decreasing standard examination time has a positive effect overall. In this case, the introduction of a linear penalty fee has the same effect as a decrease in deviation costs ν^h. Consequently, a possibly negative effect caused by the increase of ν^d (which prevails in case of $\Phi < \Phi_d$) could be outweighed by the positive effect of decreasing the costs of hastening prosecution. Whenever this is the case, a linear penalty fee—that is, an increase in ν^d and a simultaneous decrease in ν^h—has a positive welfare effect. We show in Appendix C that this is the case whenever the costs of prolonging pendency are higher than the costs of hastening it ($\nu^d > \nu^h$). In case of high costs of hypercongestion ($\Phi \geq \Phi_d$), increasing ν^d and ν^h has a positive effect—that is, a linear penalty fee has a positive welfare effect.

Summarizing the welfare effects of this policy response we state:

Result 2 *The effectiveness of introducing penalty fees depends on the costs of hypercongestion.*

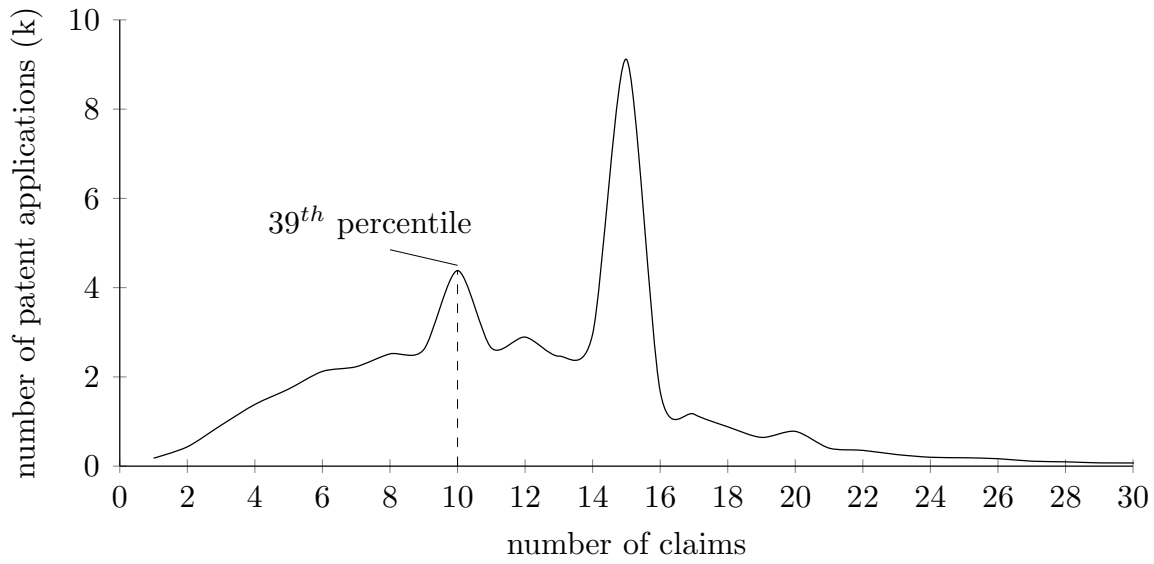
- *If the costs of hypercongestion are low, a stepped penalty fee decreases welfare, whereas a linear penalty fee increases welfare if and only if the penalty is sufficiently severe and delaying is more costly than hastening examination.*

- *If the costs of hypercongestion are high, a stepped penalty fee increases welfare, whereas a linear penalty fee increases welfare if and only if the penalty is sufficiently severe.*

Patent offices have already adopted stepped penalty fees, mainly in the form of claim-based fees. However, claim-based fees at the USPTO are probably too low to deter harmful behavior. As of April 1, 2015 inventors have to pay $80 for each claim in excess of 20. Claim-based fees at the EPO are more likely to affect behavior. They amount to €235 for each claim in excess of 15 and €580 for each claim in excess of 50. In addition, inventors must pay €15 for each page

in excess of 35 (and an additional €15 for each page in excess of 35 when the patent application is granted).

The threshold number of claims is currently set arbitrarily by patent offices. One could use an office-specific estimate of σ^{low} to determine the threshold number at which penalty fees should be set. For instance, assuming that σ^{low} at the EPO is 0.39 (see Section 5.2), the office should charge penalty fees for each claim in excess of 10. This figure corresponds to the 39th percentile of claims at the EPO as illustrated in Figure 5.



Notes: Figure truncated above 30 claims. Authors' computation based on PATSTAT data for the cohort of granted patents filed at the EPO in year 2010.

Figure 5: Distribution of the number of claims for patent applications in year 2010 at the EPO

5.5 Response 3: Alter the value of pending patents

It is possible to affect social welfare by altering the costs and benefits of pendency. Increasing the benefits of waiting in the queue is generally seen as desirable in the traffic congestion context. Think of how the spread of mobile phones has increased the productivity of commuters stuck in traffic—thereby reducing the overall cost of traffic congestion.

In our model, the value of pending patents could be altered by changing the cost of excess examination time, α. A decrease in α reduces the costs of excess examination for all inventors. However, average excess examination costs, β, are left unaffected as they are independent of α (see equation 22). Consequently changing the cost of excess examination time has no effect on welfare.

However, there is an alternative way to think about increasing the value of pending patent applications. As the value of longer pendency rises, inventors naturally desire longer durations of pendency—that is, a higher $\overline{t}_p$ and possibly also a higher $\underline{t}_p$ could result.

Increasing the patience of inventors preferring long pendencies

From equation (1) we know that $\overline{N} = s(\overline{t}_p - \underline{t}_p)$, meaning that increasing $\overline{t}_p$ while holding the number of patent applications constant is possible only if the flow of patent applications reaching their standard examination time, s, decreases. Thus, the broader spread caused by a higher $\overline{t}_p$ leads to a slower increase of the number of patent applications facing excess examination time and then to a slower decrease. Using equation (18), it is straightforward to show that a higher $\overline{t}_p$ causes the first patent application to face excess examination time earlier, $\partial \underline{t}/\partial \overline{t}_p < 0$, whereas the date at which the backlog vanishes moves further into the future, $\partial \overline{t}/\partial \overline{t}_p > 0$ (equation 19). The larger overall examination duration in turn leads to higher deviation costs ($\partial\gamma/\partial s < 0$) and possibly—depending on the relative increases in $\overline{t}$ and $\overline{t}_p$—also to an increase in the additional examination time needed by the patent office. Thus, an increase in $\overline{t}_p$ is detrimental to welfare.

Increasing the patience of all inventors

If, however, all inventors become more patient so that $\overline{t}_p$ and $\underline{t}_p$ increase, whereas $\overline{t}_p - \underline{t}_p$ remains constant, then both dates move into the future. Thus the time interval during which the cohort causes excess examination time is shifted into the future, without exerting any further effect on welfare. Consequently, if longer pendencies become more attractive and make all inventors more patient, welfare is left unaffected. Summarizing, we have

Result 3 *Increasing the value of pending patent applications (i) does not affect welfare if it decreases the costs of excess examination time; (ii) decreases welfare if it increases the patience of inventors preferring long pendencies only.*

This result suggests that it may be desirable to decrease the value of pending patents. To achieve this—that is, to increase costs of excess examination time—the patent office can seek legislative changes. However, we need to carefully distinguish between measures that affect the costs of excess examination time and those that affect the patience of inventors. To increase the costs of excess examination time, the patent office could seek charging patent pending fees. Such fees make pendency more costly for all inventors but, due to the independence of β from the cost parameter α, this policy response does not affect welfare. The EPO charges such fees

in the form of renewal fees, which are due from the third year after filing, even if the patent is not yet granted. The USPTO does not charge such fees.

A way of decreasing the patience of inventors is to systematically publish (that is, disclose) patent applications 18 months after the filing date. This practice is the norm at the EPO, but not at the USPTO (at least for domestic applications). In light of our welfare implications, publishing pending applications would lead to a positive welfare effect as it could decrease the patience of inventors seeking a long pendency. The model also provides insights on "provisional rights," which allow a patentee to backdate damages to include infringements that occurred prior to the grant of a patent. Provisional rights reduce the opportunity cost faced by inventors, which increases their patience and therefore harms welfare.

6 Concluding remarks

This paper provides a model of the backlog at patent offices. It advances knowledge in three main ways.

First, it connects the urban economics literature with the economics of innovation literature. It adapts the dynamic bottleneck model of traffic congestion to the patent system, thereby allowing congestion pricing studies in this new context. A handful of scholars have established a parallel between traffic congestion and patent backlogs, but this paper is the first to propose a full-fledged model of the sort.

Second, the paper brings into sharp focus the interaction between filing strategy and pendency. Patent attorneys are well aware that drafting style affects pendency, but this fact has largely escaped economists' attention. We find that the backlog impedes the progress of patent examination by providing incentives to strategically manipulate pendency.

Third, the analysis comes with a series of implications that are particularly relevant from a policy viewpoint. Contrary to received wisdom, the model suggests that increasing the patent office's processing capacity is not necessarily an appropriate response given the apparent low proportion of low-workload applications. The model also suggests that a penalty fee that would hit *all* applications can be more effective than a fee that would predominantly target high-workload applications. The intuition behind this result is that a linear penalty fee makes low-workload filing relatively less costly and hence more attractive. The result holds even if the penalty fee targets backlog-inducing characteristics such as the number of claims, pages, and communications. In other words, the model suggests that a fee that increases linearly with the number of claims has a greater welfare effect than a fee applied to claims above a certain

threshold. If the actual costs of hypercongestion are high enough, either fee enhances welfare. However, linear penalty fees must be sufficiently high to lead to welfare improvement. Finally, increasing the value of pending applications is not generally welfare enhancing. Instead, patent offices should consider reforms that decrease the value of pendency, such as renewal fees for pending applications, higher fees for continued examination, more restrictive provisional rights, and systematic disclosure of applications after 18 months.

A broader implication from the study derives from the model's focus on drafting style. One root cause of "probabilistic" patents in the sense of Lemley and Shapiro (2005) is a deliberate act by inventors to draft vague descriptions in an attempt to extend the scope of their rights (see, for example O'Neill et al., 2007). Scholars have shown that such patents are socially harmful (Choi, 2005; Farrell and Shapiro, 2008). Devising policy instruments that induce inventors to improve drafting quality helps not only to reduce the backlog but also to clarify the validity of intellectual property rights.

We see three avenues for future research. First, one could build on our model to study other policy-relevant questions such as the patent prosecution highway initiative and the practices of deferred examination and fast-tracking of applications.[10] Deferred examination gives inventors preferring long pendency the opportunity to postpone examination to a later date. In our model, this practice would amount to opening up a second lane where inventors could park for a while. Such a practice temporarily reduces the size of the cohort, leaving the office more resources to work on applications from inventors who are in a hurry to have their patent granted. However, because it does little to solve the technological and market uncertainty associated with pendency, its overall welfare effect is unclear.

Second, we know very little about how fast applicants want their patent applications to be granted (the parameter σ^{low} in the model). Yet, this parameter is key to designing a patent system that best balances the needs of inventors with the needs of society. Large-scale surveys of patent applicants could collect such information and enlighten economists and policy analysts. The EPO already surveys applicants for budget-planning purposes, and it would be straightforward to include additional questions.

Third, testing the model with data will further deepen our understanding of the economics of backlog. Policy reforms such as the introduction of a penalty fee or a sudden change in examination capacity provide ideal natural experiments to improve our understanding of the issues discussed in this paper. We hope that our model provides a useful starting point for all

[10]The PPH aims to accelerate the prosecution of international patent applications by improving information sharing between offices.

these research questions.

References

Acemoglu, Daron, and Ufuk Akcigit. 2012. "Intellectual property rights policy, competition and innovation." *Journal of the European Economic Association*, 10(1): 1–42.

Arnott, Richard, André de Palma, and Robin Lindsey. 1990. "Economics of a bottleneck." *Journal of Urban Economics*, 27(1): 111–130.

Arnott, Richard, André de Palma, and Robin Lindsey. 1993. "A structural model of peak-period congestion: A traffic bottleneck with elastic demand." *American Economic Review*, 83(1): 161–179.

Berger, Florian, Knut Blind, and Nikolaus Thumm. 2012. "Filing behaviour regarding essential patents in industry standards." *Research Policy*, 41(1): 210–225.

Choi, Jay P. 2005. "Live and let live: A tale of weak patents." *Journal of the European Economic Association*, 3(2-3): 724–733.

Conti, Annamaria, Jerry Thursby, and Marie C. Thursby. 2013. "Patents as signals for startup financing." *The Journal of Industrial Economics*, 61(3): 592–622.

DeBrock, Lawrence M. 1985. "Market structure, innovation, and optimal patent life." *Journal of Law and Economics*, 28(1): 223–244.

de Rassenfosse, Gaétan, Alfons Palangkaraya, and Elizabeth Webster. 2016. "Why do patents facilitate trade in technology? Testing the disclosure and appropriation effects." *Research Policy*, 45(7): 1326–1336.

de Rassenfosse, Gaétan, and Timo Fischer. 2016. "Venture debt financing: Determinants of the lending decision." *Strategic Entrepreneurship Journal*, forthcoming.

Farrell, Joseph, and Carl Shapiro. 2008. "How strong are weak patents?" *American Economic Review*, 98(4): 1347–1369.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist. 2015. "The bright side of patents." *USPTO Economic Working Paper No. 2015-5.*

Gallini, Nancy. 2002. "The economics of patents: Lessons from recent US patent reform." *The Journal of Economic Perspectives*, 16(2): 131–154.

Gans, Joshua S., David H. Hsu, and Scott Stern. 2008. "The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays." *Management Science*, 54(5): 982–997.

Graham, Stuart J. H. 2006. "The determinants of patentees' use of 'continuation' patent applications in the United States Patent and Trademark Office 1980–99." In *Intellectual property rights: Innovation, governance and the institutional environment.* , ed. B. Andersen, Chapter 7, 215–241. Cheltenham and Northampton: Edward Elgar.

Graham, Stuart J. H., and Galen Hancock. 2014. "The USPTO economics research agenda." *The Journal of Technology Transfer*, 39(3): 335–344.

Harhoff, Dietmar. 2016. "Patent Quality and Examination in Europe." *American Economic Review Papers and Proceedings*, 106(5): 193–197.

Harhoff, Dietmar, and Stefan Wagner. 2009. "The duration of patent examination at the European Patent Office." *Management Science*, 55(12): 1969–1984.

IP5 Offices. 2015. "IP 5 Statistics Report 2014 Edition." *European Patent Office, Munich.*

Koenen, Johannes, and Martin Peitz. 2012. "The economics of pending patents." In *Recent advances in the analysis of competition policy and regulation.* , ed. J. Harrington and Y. Katsoulacos, Chapter 3. Cheltenham and Northampton: Edward Elgar.

Lazaridis, George, and Bruno van Pottelsberghe de la Potterie. 2007. "The rigour of EPO's patentability criteria: An insight into the 'induced withdrawals"." *World Patent Information*, 29(4): 317–326.

Lemley, Mark. A., and Carl Shapiro. 2005. "Probabilistic patents." *Journal of Economic Perspectives*, 19(2): 75–98.

London Economics. 2010. "Patent backlogs and mutual recognition." *UK IPO, London.*

Mabey, Warren K. 2010. "Deconstructing the Patent Application Backlog." *Journal of the Patent and Trademark Office Society*, 92(2): 208–283.

Marco, Alan C., and James E. Prieger. 2009. "Congestion pricing for patent applications." *http://dx.doi.org/10.2139/ssrn.1443470.*

Matutes, Carmen, Pierre Régibeau, and Katharine Rockett. 1996. "Optimal patent design and the diffusion of innovations." *The RAND Journal of Economics*, 27(1): 60–83.

Mitra-Kahn, Benjamin, Alan Marco, Michael Carley, Paul DAgostino, Peter Evans, Carl Frey, and Nadiya Sultan. 2013. "Patent backlogs, inventories, and pendency: An international framework." *UK Intellectual Property Office Working Draft 2013/25.*

O'Neill, Sean, Kirk Hermann, Marlene Klein, Jeff Landes, and Raj Bawa. 2007. "Broad claiming in nanotechnology patents: Is litigation inevitable." *Nanotechnology Law & Business*, 4(1): 595–606.

Osenga, Kristen J. 2005. "Entrance ramps, tolls, and express lanes – Proposals for decreasing traffic congestion in the patent office." *Florida State University Law Review*, 33(2): 119–156.

Palangkaraya, Alfons, Paul H Jensen, and Elizabeth Webster. 2008. "Applicant behaviour in patent examination request lags." *Economics Letters*, 101(3): 243–245.

Picard, Pierre M., and Bruno van Pottelsberghe de la Potterie. 2013. "Patent office governance and patent examination quality." *Journal of Public Economics*, 104(August): 14–25.

Popp, David, Ted Juhl, and Daniel Johnson. 2004. "Time in purgatory: Examining the grant lag for US patent applications." *The B. E. Journal of Economic Analysis & Policy*, 4(1): 1–45.

Reitzig, Markus, Joachim Henkel, and Christopher Heath. 2007. "On sharks, trolls, and their patent prey — Unrealistic damage awards and firms' strategies of 'being infringed'." *Research Policy*, 36(1): 134–154.

Sharon, Ayal, and Yifan Liu. 2007. "Improving patent examination efficiency and quality: An operations research analysis of the USPTO, using queuing theory." *Federal Circuit Bar Journal*, 17(2): 133–164.

Small, Kenneth A., and Erik T. Verhoef. 2007. "The Economics of Urban Transportation." *London and New York: Routledge.*

Sperber, Philip. 1970. "The strategy of delaying and expediting patent prosecution." *Journal of the Patent Office Society*, 52(3): 141–177.

Vickrey, William S. 1969. "Congestion theory and transport investment." *American Economic Review Papers and Proceedings*, 59(2): 251–261.

A Derivation of equilibrium values

To derive an equilibrium we identify the filing strategy for which the cutoff inventor is indifferent between choosing either the lowest-workload, the highest-workload, or her optimal filing strategy. The lowest-workload filing strategy is given by $t_\delta = \underline{t}$ and the highest-workload filing strategy by $t_\delta = \overline{t}$. In both extreme cases excess examination time is zero. For $\underline{t} = 0$, this is so because the patent is "first in line"; for $\overline{t}$ it is because it is "last in line." Thus with either the lowest- or the highest-workload filing strategy the cutoff inventor faces only deviation costs. Indifference of the cutoff inventor constitutes an equilibrium because inventors with a marginally lower (higher) preferred duration of pendency always choose a lower (higher) standard examination time to minimize deviation costs. Consequently, given that the cutoff inventor is indifferent, the best filing strategies for all other inventors are: choose a lower-workload filing strategy than the cutoff inventor for $t_p < \hat{t}_p$ or a higher-workload filing strategy than the cutoff inventor for $t_p > \hat{t}_p$.

The cutoff inventor has no incentive to deviate if and only if deviating does not outperform not deviating. When choosing the lowest-workload filing strategy the cutoff inventor would incur deviation costs of $\nu^h(\hat{t}_p - \underline{t})$. When choosing the highest-workload filing strategy the cutoff inventor would face deviation costs of $\nu^d(\overline{t} - \hat{t}_p)$. Thus she has to trade off maximal excess examination costs against maximal deviation costs. This yields equilibrium condition

$$\underbrace{\nu^h(\hat{t}_p - \underline{t})}_{(I)} = \underbrace{\alpha(\hat{t}_p - \hat{t}_\delta)}_{(II)} = \underbrace{\nu^d(\overline{t} - \hat{t}_p)}_{(III)}. \tag{A.1}$$

Equating (I) and (III) and using $\overline{t} = s(\overline{t}_p - \underline{t}_p)/\kappa + \underline{t}$ from equation (13), we can compute $\hat{t}_p$ as

$$\hat{t}_p = \underline{t} + \frac{\nu^d \overline{N}}{(\nu^h + \nu^d)\kappa}. \tag{A.2}$$

where $\overline{N} = s(\overline{t}_p - \underline{t}_p)$ (equation 1).

The excess examination time that the cutoff inventor with desired pendency $\hat{t}_p$ faces in equilibrium, $\hat{X}(t_\delta)$, can be calculated by inserting equation (A.2) into the equilibrium condition $(I) = (II)$ using the relation $\hat{X}(t_\delta) = \hat{t}_p - \hat{t}_\delta$,

$$\hat{X}(t_\delta) = \frac{\nu^d \nu^h \overline{N}}{(\nu^h + \nu^d)\alpha\kappa}. \tag{A.3}$$

From equation (15) we know that the share of low-workload filing strategies is given by $\sigma^{low}\overline{N} = \alpha\kappa(\hat{t}_\delta - \underline{t})/(\alpha - \nu^h)$. Inserting the relation $\hat{t}_\delta = \hat{t}_p - \hat{X}(t_\delta)$ and plugging in $\hat{t}_p$ from equation (A.2) as well as $\hat{X}_\delta$ from equation (A.3) gives us the equilibrium value of σ^{low}

$$\sigma^{low} = \frac{\nu^d}{\nu^d + \nu^h}. \tag{A.4}$$

Solving equation (14) for $\hat{t}_p$ and plugging this quantity into equation (A.2) yields $\underline{t} = \sigma^{low}(\overline{t}_p - \underline{t}_p) + \underline{t}_p - \sigma^{low}\overline{N}_e/\kappa$. Inserting the relation $\underline{t}_p = \overline{t}_p - (\overline{t}_p - \underline{t}_p)$ gives us

$$\overline{t} = \overline{t}_p + (1 - \sigma^{low})(\overline{t}_p - \underline{t}_p)\left[\frac{s}{\kappa} - 1\right] \tag{A.5}$$

as the date when the backlog vanishes. Finally, using the relation $\overline{t} = \underline{t} + \overline{N}/\kappa$, we can calculate the date when the first patent faces excess examination time as

$$\underline{t} = \underline{t}_p - \sigma^{low}(\overline{t}_p - \underline{t}_p)\left[\frac{s}{\kappa} - 1\right]. \tag{A.6}$$

B Experts' opinion on σ^{low}

We contacted twelve current and former chief economists from seven patent offices (CIPO, EPO, IP Australia, JPO, UKIPO, USPTO, and WIPO) to ask them the following question:

> "In your opinion:
> 1. What is the proportion of patent applications for which applicants intentionally delayed the decision regarding the issuance?
> 2. What is the proportion of patent applications for which applicants intentionally expedited the decision regarding the issuance?
> 3. What is the proportion of patent applications for which applicants did not intentionally affect the speed of issuance?
> In all cases think of patent applications filed in the recent years at the office for which you work (or used to work). Answers to questions 1–3 should sum up to one."

We guaranteed anonymity of the respondents because of the sensitive nature of the information provided. Seven experts answered the question: two experts declined to answer for lack of knowledge; and three experts did not reply after two reminder emails.

C Policy responses: Introducing penalty fees

For a linear penalty fee the positive welfare effect of decreasing ν^h overcompensates the negative welfare effect (where $\Phi < \Phi_d$) of increasing ν^d whenever

$$\left|\frac{\partial \Omega}{\partial \nu^h}\right| > \left|\frac{\partial \Omega}{\partial \nu^d}\right|.$$

Both derivatives consist of the following parts, which we will compare subsequently

$$\frac{\partial \Omega}{\partial \nu^\tau} = \underbrace{\frac{\partial \sigma^{low}}{\partial \nu^\tau}\Phi}_{(a)} \underbrace{-w(\kappa)\frac{\partial(\overline{t}-\overline{t}_p)}{\partial \nu^\tau}}_{(b)} \underbrace{-\frac{\partial \beta}{\partial \nu^\tau}\Phi}_{(c)} \underbrace{-\frac{\partial \gamma}{\partial \nu^\tau}}_{(d)}, \tag{C.1}$$

where $\tau = d, h$.

Part (a)

Using the equilibrium value for σ^{low} from equation (16) the partial derivatives are given by $\frac{\partial \sigma^{low}}{\partial \nu^h} = \frac{-\nu^d}{(\nu^d+\nu^h)^2}$ and $\frac{\partial \sigma^{low}}{\partial \nu^d} = \frac{\nu^h}{(\nu^d+\nu^h)^2}$. Comparing the absolute values of part (a) for $\tau = h$ and $\tau = d$ respectively, we have $\left|\frac{\partial \sigma^{low}}{\partial \nu^h}\Phi\right| > \left|\frac{\partial \sigma^{low}}{\partial \nu^d}\Phi\right|$ if $\nu^d > \nu^h$.

Part (b)

Using the equilibrium value for $\overline{t}-\overline{t}_p$ from equation (19) yields $\frac{\partial(\overline{t}-\overline{t}_p)}{\partial \nu^\tau} = -\frac{\partial \sigma^{low}}{\partial \nu^\tau}(\overline{t}_p - \underline{t}_p)\left[\frac{s}{\kappa} - 1\right]$. Thus, as before, the comparison of the absolute values of part (b) for $\tau = h$ and $\tau = d$ depends on the partial derivatives of σ^{low} with respect to deviation costs. Given the above result of $\left|\frac{\partial \sigma^{low}}{\partial \nu^h}\right| > \left|\frac{\partial \sigma^{low}}{\partial \nu^d}\right|$ if $\nu^d > \nu^h$ we have $\left|-w(\kappa)\frac{\partial(\overline{t}-\overline{t}_p)}{\partial \nu^h}\right| > \left|-w(\kappa)\frac{\partial(\overline{t}-\overline{t}_p)}{\partial \nu^d}\right|$ if $\nu^d > \nu^h$.

Part (c)

The partial derivatives of the average costs of excess examination, β (equation 22), are given by $\frac{\partial \beta}{\partial \nu^d} = \frac{\nu^{2h}\overline{N}}{(\nu^l+\nu^s)^2 2\kappa}$ and $\frac{\partial \beta}{\partial \nu^h} = \frac{(\sigma^{low})^2}{2\kappa}$. Comparing the absolute values of part (c) yields $\left|\frac{\partial \beta}{\partial \nu^h}\right| > \left|\frac{\partial \beta}{\partial \nu^d}\right|$ if $\nu^d > \nu^h$.

Part (d)

The partial derivatives of the average deviation costs, γ (equations 23 and 24), are given by $\frac{\partial \gamma}{\partial \nu^d} = \frac{\nu^{2h}\overline{N}}{(\nu^d+\nu^h)^2 2\kappa}\left[1 - \frac{\kappa}{s}\right]$ and $\frac{\partial \gamma}{\partial \nu^h} = \frac{\sigma^{low})^2\overline{N}}{2\kappa}\left[1 - \frac{\kappa}{s}\right]$. Again, comparing absolute values yields $\left|\frac{\partial \gamma}{\partial \nu^h}\right| > \left|\frac{\partial \gamma}{\partial \nu^d}\right|$ if $\nu^d > \nu^h$.

For all parts of equation (C.1) the absolute values of the welfare effects of decreasing the costs of shortening pendency, ν^h, are higher than the absolute values of increasing the costs of prolonging pendency as long as $\nu^d > \nu^h$, meaning that in this case the positive effect of a linear penalty fee outweighs the negative effect.