CEEPR Center for Energy and Environmental Policy Research

Testing the Efficiency of a Tradeable Permits Market

by
Juan-Pablo Montero

02-004 WP September 2002

A Joint Center of the Department of Economics, Laboratory for Energy and the Environment, and Sloan School of Management

Testing the Efficiency of a Tradeable Permits Market

Juan-Pablo Montero*

Catholic University of Chile and MIT

September 6, 2002

Abstract

A tradeable permits market is said to be efficient when all affected firms trade permits until their marginal costs equal the market price. Detailed firm-level data are generally required to perform such an efficiency test, yet such information is rarely available. If firms face a declining target, however, and are allowed to bank permits, as has occured recently, aggregated data such as the evolution of the permits bank is sufficient to test for either less than optimal market participation or the exercise of market power. An application to the U.S. sulfur dioxide emission permits market is provided.

*<jpmonter@mit.edu> Professor of Economics at the Catholic University of Chile (PUC) and Research Associate at the MIT Center for Energy and Environmental Policy Research (CEEPR). This paper was written while I was a Visiting Professor of Applied Economics at the MIT Sloan School of Management. I am grateful to Denny Ellerman, Paul Joskow, Matti Liski, and Dick Schmalensee for helpful discussions, and to CEEPR and the U.S. EPA (STAR grant award #R-82863001-0) for financial support.

1 Introduction

It is well known that in theory, a tradeable permits program can ration a given quantity of a resource (e.g., clean air, lead in gasoline, fish, water, bus licenses, taxi medallions) at the least cost to society. The argument rests on the assumption that an *efficient* permits market will develop in which *all* affected parties trade permits until their marginal costs equal the market permits price. Even if an active market for permits develops, however, when some firms either do not fully participate in the market (due to, for example, significant transaction costs, regulatory rulings, or information asymmetries) or able to exercise market power, this active market will fail to deliver the least-cost solution.[1] And because cost information at the firm level is generally limited, as in any other market, it will be difficult to test whether or not the permits market is actually delivering the least-cost solution.

A salient example is the U.S. Acid Rain Program, for which recent studies have reached opposing conclusions regarding the performance of its sulfur dioxide (SO_2) permits market.[2] On the one hand, based on price and quantity data from private transactions, Joskow et al. (1998) argued that by mid-1994 the SO_2 market has become reasonably "efficient" in the sense that there was a large and increasing volume of transactions taking place at a single price.[3] This trading activity was consistent with the significant cost savings estimated by Ellerman et al. (2000). On the other hand, Carlson et al. (2000) constructed econometric abatement cost estimates for each individual firm in the program and used them to predict the outcome of an efficient market for the years 1995 and 1996. They found that the actual cost of compliance was not only more than 50% larger than the cost of their efficient market prediction, but it was also larger than the cost under no trading. They attributed these results to many firms' reluctance to

[1] An active market here is one characterized by significant trading activity and no arbitrage opportunities. If no such market develops, total costs will be higher than the least-cost solution, by definition; except in extremely unlikely situations in which firms are homogeneous or the regulator has sufficient information to allocate permits in the least-cost manner.

[2] The U.S. Acid Rain program calls for a 50% nationwide reduction in electric utilities' SO_2 emissions. In this particular program, permits are called allowances. See Ellerman et al. (2000) for more details.

[3] Note that, because the authors restricted their analysis to trading activity, their definition of "efficiency" is narrower than the definition used in this paper.

fully participate in the market because of poor incentives provided by existing utility price regulation.[4]

Clearly, the above findings invite further investigation not only of the SO_2 program, but more generally, of alternative ways to test the efficiency of a tradeable permits market. In this regard, an interesting aspect of the SO_2 program has not yet been explored that could shed light on this efficiency issue: as a result of a declining SO_2 permits cap firms, have engaged not only in so-called spatial trading (trading between different entities within each period) but also in intertemporal trading, i.e., they are banking permits for future use. Because the evolution of a permits bank is closely related to the evolution of an exhaustible resource stock,[5] in this paper I draw upon both the literature on tradeable permits markets and the literature on exhaustible resources to ask whether the evolution of readily available (aggregate) data such as prices and the permits bank provide enough information to detect less than optimal market participation or the exercise of market power.[6]

Although my motivation derives from the SO_2 market, the model developed in this paper and its implications apply more broadly to any tradeable permits market that faces a declining target and allows for banking, a possibility that is attracting attention as a way to gradually introduce new regulation or tighten existing regulation.[7] The rest of the paper is organized as follows. Section 2 presents the model and examines the performance of a permits market in which, for either regulatory or economic reasons, firms

[4] Although not mentioned by the authors, higher compliance costs can also be due, at least in principle, to market power.

[5] Important differences exist, though. First, the permits market remains after the permits bank has been exhausted, while the market for a typical exhaustible resource vanishes after the total stock has been consumed. Second, storage costs for permits are zero, while they are generally positive for a typical exhaustible resource. In addition, the demand for permits corresponds to a derived demand from the same firms that hold the permits, while the demand for a typical exhaustible resource comes from a third party.

[6] Note that I explicitly say "optimal market participation" instead of "broad participation," because, in principle, the permits allocation may be such that little trading is required to reach the least-cost solution.

[7] President Bush's Clear Skies Initiative would reduce the existing SO_2 emissions cap by another 70% in two steps starting in 2010 and current legislative proposals before the U.S. Congress would effect similar reductions. Moreover, an effective policy for reducing atmospheric greenhouse gas concentrations would likely include emission caps that would become more stringent over time. A banking provision was also included as part of the permits program that gradually phased down the use of lead in gasoline during the 1980s.

do not bank permits but only trade permits spatially. Section 3 introduces a declining target and banking, and explores the effect of non-participation in the market on the evolution of aggregate data such as the permits bank. Section 4 analyzes the evolution of prices and the permits bank for a dominant firm with a competitive fringe. Section 5 applies the model to SO_2 market data. Final remarks are offered in Section 6.

2 The model

For concreteness and consistency with the application in Section 5, the model that developed here focuses on an emission permits market, but it can easily be extended to any other permits market by simply relabeling the variables.

2.1 Variables

Consider an industry with a large number N of heterogeneous firms whose emissions are regulated by a tradeable permits program (in the study of market power, I will assume that a large fraction of these firms merge to become a von-Stackelberg-dominant producer). The regulator allocates a total of $A(t) = \sum_{i=1}^{N} a_i(t)$ allowances (or permits) in period t, where $a_i(t)$ is firm i's allocation at t (note that capital letter will denote industry or group-level variables and small letter will denote firm-level variables). Individual and aggregate allocations, which can vary over time, are common knowledge.

Firms differ in their costs of abatement and unrestricted emissions (i.e., emissions that would have been observed in the absence of the permits program). For mathematical tractability, I assume that firm i's abatement costs are given by

$$c_i(q_i(t)) = \alpha_i [q_i(t)]^{\beta+1} \tag{1}$$

where $q_i(t)$ are emissions reduced at t, $\alpha_i > 0$ is firm i's private information, and $\beta \geq 1$ (to ensure an interior solution) is common to all firms and known to the regulator.[8]

[8]Note that letting β vary across firms adds generality to the model but makes it mathematically untractable. I discuss the implications of relaxing this assumption in a few places later in the paper and argue that results are not affected by this assumption. Also, we can let α_i decrease over time

For a given aggregate level of reduction $Q(t)$, the industry least-cost reduction burden solves $c_i'(q_i(t)) = c_j'(q_j(t))$ for all $i \neq j$. Letting $Q(t) = \sum q_i(t)$, the industry least-cost curve becomes

$$C(Q(t)) = \gamma[Q(t)]^{\beta+1} \tag{2}$$

where

$$\gamma = \left(\sum_{i=1}^{N} \left(\frac{1}{\alpha_i} \right)^{1/\beta} \right)^{-\beta}$$

As commonly assumed in the literature (e.g., Weitzman, 1974), this cost formulation supposes that the regulator has some notion about the shape of the aggregate cost function, but not enough to predict the market equilibrium price for a given level of aggregate reduction.

Firm i's unrestricted or counterfactual emissions are denoted by $u_i(t)$. Thus, firm i's emissions at time t are $e_i(t) = u_i(t) - q_i(t)$, and industry-level emissions are $E(t) = \sum_{i=1}^{N} e_i(t)$. As with costs, I assume that the regulator (or analyst) has much better information at the aggregate than at the individual level (or that estimations at the aggregate level are more precise than at the individual level), so I assume that he knows $U(t) = \sum_{i=1}^{n} u_i(t)$ but not $u_i(t)$.[9]

The above heterogeneity in costs and unrestricted emissions assures that many firms must buy and sell permits in the market in order to minimize total compliance costs. Firm i's trading volume in period t is $x_i(t)$, which can be either positive or negative depending on whether the firm is a net buyer ($x_i > 0$) or net seller ($x_i < 0$) of permits. Because the regulator directly observes firm i's emissions e_i (information that is also available to the analyst) and enforces full compliance, x_i is known and equal to $e_i - a_i$. Note that I simply define trading volume as the difference between emissions and allowance allocation, regardless of the number of market exchanges in which the firm was actually

at some known industry-wide rate $\delta \geq 0$ as a result of an exogenous innovation trend, specifically $\alpha_i(t) = \alpha_i(0)\mathrm{e}^{-\delta t}$, without any change in the resolution of the model.

[9] I discuss this assumption further in the numerical section.

engaged. I am solely interested in the firm's final position in the market, whether as a net buyer or net seller.

2.2 A market without banking

Before introducing banking, let us consider the simplest case, in which the regulator does not allow firms to bank permits for future use (alternatively, consider the allowance cap $A(t)$ to be constant over time so there are no incentives for banking). When there is no banking, full compliance implies $A(t) = E(t)$ for all t, so market efficiency, as defined above, requires (Montgomery, 1972)

$$C'(Q(t)) = P^*(t) \tag{3}$$

for all t, where $Q(t) = U(t) - A(t)$, $C'(Q(t))$ is obtained from (2), and $P^*(t)$ denotes the "optimal" price of permits. Because γ is not known with precision, however, we cannot be certain whether the observed market price $P^\circ(t)$ is the efficient price or not (hereafter the superscript "$\circ$" will indicate the observed variable). Even if γ is known, $P^\circ(t) = P^*(t)$ is not a sufficient condition for market efficiency. Although we can rule out the presence of market power, we can still have $P^\circ = P^*$ with less than optimal market participation.

Market efficiency also requires a certain volume of trading. If $x_i(t)$ is the number of permits that firm i trades during period t, full compliance requires $x_i(t) + a_i(t) = u_i(t) - q_i(t)$. Thus, in an efficient market, firm i's trading volume is

$$x_i(t) = u_i(t) - a_i(t) - \left(\frac{P^*(t)}{\alpha_i}\right)^{1/\beta}$$

Replacing $P^*(t)$ according to (2) and (3), the (optimal) individual and aggregate volumes of trading become, respectively,

$$x_i^*(t) = u_i(t) - a_i(t) - \left(\frac{\gamma}{\alpha_i}\right)^{1/\beta} Q(t) \tag{4}$$

$$V^*(t) = \sum_{i=1}^{N} \left| u_i(t) - a_i(t) - \left(\frac{\gamma}{\alpha_i}\right)^{1/\beta} Q(t) \right| \tag{5}$$

where $Q(t) = U(t) - A(t)$.

Unlike price information, the actual or observed trading volume $V^\circ(t)$ does convey enough information for the analyst to conclude whether or not the market is efficient. In fact, $V^\circ(t) < V^*(t)$ whenever there is non-participation (Stavins, 1995; Montero, 1998) or market power (Hahn, 1984). However, when u_i, γ/α_i or both of these values are unknown, neither x_i^* nor V^* can be estimated with precision;[10] hence, the observed trading volume (whether at the aggregate level or the individual level) does not tell us per se whether or not the market is efficient. While a market with a significant trading volume and broad participation is more likely to approach efficiency, a market with a relatively low trading activity, where $x_i^* \approx 0$ for several firms, cannot be ruled as inefficient.

Thus, in the absence of detailed individual-level data on costs and unrestricted emissions, it is not possible to conclude from trading activity data whether or not the market is delivering the least-cost solution. Because it is always difficult to collect and develop accurate firm-level information,[11] looking at each firm's final position in the market (i.e., x_i°) and comparing this to estimates of x_i^* and V^* becomes almost a futile exercise. The next two sections tackle the same efficiency question using a different approach that relies on aggregate information and the intertemporal properties of an efficient equilibrium.

3 A market with banking

Let us consider the same N firms, but in a market in which the regulator allows firms to bank permits for future use. For the latter to actually happen, permits allocations must decrease over time (at least at a rate higher than the discount rate for some period of

[10] Note that because $a_i(t)$ varies across firms, $V^*(t)$ cannot be estimated even if the correlation between u_i and α_i is known.

[11] As discussed by Ellerman et al. (2000, Appendix) for the case of unrestricted emissions, econometric methods can provide reasonably accurate estimations for aggregate variables but generally imprecise estimates for individual variables (i.e., α_i and u_i). Accounting for individual statistical errors in the estimation of V^* would lead to such a wide confidence interval that contrasting V° to V^* would be of little use.

time). To simplify notation and follow the SO_2 program design, I let a_H be the (high) per-period permits allocation of each firm during the first T periods of the program and a_L be the (low) per-period allocation thereafter, with $a_H \gg a_L$. Thus, the aggregate allocations during these two phases are $A_H = a_H N$ and $A_L = a_L N$, respectively.[12] To simplify notation further without loss of generality, I assume that $u_i(t)$ remains constant over time, so I drop its time index.

3.1 The efficient solution

As in the static case, an efficient market with banking solves the following infinite horizon intertemporal minimization problem (Rubin, 1996; Schennach, 2000)

$$\min \int_0^\infty \left(\sum_{i=1}^{N} c_i(q_i(t)) \right) \mathrm{e}^{-rt} dt \tag{6}$$

$$\text{s.t. } \dot{B}(t) = A(t) + Q(t) - U \tag{7}$$

$$B(0) = 0, -B(t) \leq 0 \tag{8}$$

where r is the risk-free discount rate, $B(t)$ is the stock (i.e., bank) of allowances at time t and the dot denotes a time derivative. Denoting by $\lambda(t)$ and $\phi(t)$ the multiplier functions, the Hamiltonian for this problem can be written as

$$H = C(Q(t))\mathrm{e}^{-rt} + \lambda(A(t) + Q(t) - U) - \phi B(t)$$

where $C(Q(t))$ is given by (2).

Necessary conditions for optimality include satisfaction of (7), (8), and[13]

$$\frac{\partial H}{\partial Q} = C'(Q(t))\mathrm{e}^{-rt} + \lambda(t) = 0 \tag{9}$$

$$\dot{\lambda}(t) = -\frac{\partial H}{\partial B} = \phi(t) \tag{10}$$

$$\phi(t) \geq 0, \ \ \phi(t)B(t) = 0 \tag{11}$$

[12] Alternatively, one can let $(A_H - A_L)T$ be the initial stock $B(0)$ and A_L be the annual allocation for every period.

[13] See Kamien and Schwartz (1991).

In addition, taking the derivative of (9) with respect to time yields

$$\dot{C}'(Q(t)) - rC(Q(t)) + \phi e^{rt} = 0 \tag{12}$$

When $B(t) > 0$, $\phi(t) = 0$ and marginal costs, and hence prices, follow Hotelling's rule, rising at the discount rate r (note that permits are "extracted" at zero cost).

Whether and when firms will bank permits depends upon the allocation of permits, the evolution of marginal cost functions, and the discount rate. For example, a significant reduction of the permits allocation in the future, as in the SO_2 program, will result in a banking period of some length τ (to be determined shortly): firms bank permits during some period of time and gradually use them thereafter, until the bank expires at τ. After τ, firms go back to the "market without banking" situation described above.

The full compliance condition establishes that the total number of permits allocated during the banking period $[0, \tau]$ be equal to the emissions accumulated during such period, that is[14]

$$(A_H - A_L)T + A_L\tau = \int_0^\tau [U - Q(t)]dt \tag{13}$$

At τ, the terminal condition $E(\tau) = A(\tau)$ must also hold

$$Q(\tau) = U - A_L \tag{14}$$

From (12) we have that $C'(Q(t)) = C'(Q(\tau))e^{-r(\tau - t)}$ when $\phi = 0$. Then, using (2) and (14), we have that

$$Q(t) = (U - A_L)e^{-r(\tau - t)/\beta} \tag{15}$$

Substituting (15) into (13) and rearranging, we obtain the following expression that solves

[14]Note that this condition is the exhaustion condition of the exhaustible resource literature.

for τ

$$\frac{(A_H - A_L)T}{U - A_L} = \frac{(a_H - a_L)T}{\overline{u} - a_L} = \tau - \frac{\beta}{r}\left(1 - \mathrm{e}^{-r\tau/\beta}\right) \tag{16}$$

where $\overline{u} = U/n$. Thus, for known values of A_H, A_L, U, β, and r, (16) provides a unique solution τ^*, which in turn allows us to compute the efficient paths for prices and quantities during the banking period $[0, \tau^*]$. $P^*(t)$ will increase at the interest rate starting from $P^*(0) = \gamma(\beta+1)(U - A_L)^{\beta}\mathrm{e}^{-r\tau^*}$, while $Q^*(t)$ will increase at the rate r/β, starting from $Q^*(0) = (U - A_L)\mathrm{e}^{-r\tau^*/\beta}$. From the latter we can also compute the optimal emission path, $E^*(t) = U - Q^*(t)$, and the optimal banking path, $B^*(t) = \int_0^t [A(t) + Q^*(t) - U]dt$. Because the evolutions of these two quantity variables are directly connected to $Q(t)$ in what follows, I focus on $Q(t)$.

3.2 The effect of limited market participation

Having derived the aggregate behavior of an efficient market based on either data that are readily available (e.g., A_H, A_L) or parameters that can be estimated with reasonable precision (e.g., U, β, and r),[15] one question remaining is whether we can detect less than optimal market participation by contrasting the observed paths of prices $P^\circ(t)$ and quantities $Q^\circ(t)$ with the optimal paths $P^*(t)$ and $Q^*(t)$. As in the non-banking case, P° provides little information to answer such a question, not only because of uncertainty about γ but also because arbitrageurs ensure that $\dot{P}^\circ/P^\circ = r$ in any active market. Conversely, the evolution of Q° (or B°) can provide valuable information to detect suboptimal market participation.

To study the effect of non-participation on $Q(t)$, let us define firm j as a non-participant if it uses its own permits for compliance and discounts the future at some rate r_j, which can be greater than, equal to, or lower than r.[16] In the extreme case, a firm using $r_j = \infty$ (or sufficiently large) will bank no permits, and $q_j(t) = u_j - a_i(t)$ for all t.

[15] See Ellerman and Montero (2002) for a discussion on how to collect this data for the SO_2 market.

[16] Note that this formulation is general enough to accommodate for inefficient participation rather than non-participation. Simply break down the firm into an arbitrarily large number of production units and let some fraction of these units not participate in the market.

This is a non-participating firm that even fails to minimize costs intertemporally, given its own endowment of permits. In comparing Q^* with Q° for a market with less than full participation, it is useful to split the analysis into two cases: (a) all non-participating firms discount the future at some rate other than r, and (b) all non-participating firms use r.

Consider first for case (a), without loss of generality, two non-participating firms 1 and 2 with discount rates r_1 and r_2. Since the optimization problem solved by either of these production firms is similar to the optimization problem for the industry, the reduction path followed by a non-participating firm $j = 1, 2$ is (superscript "n" stands for non-participation)

$$\frac{\dot{q}_j^n(t)}{q_j^n(t)} = \frac{r_j}{\beta} \tag{17}$$

where $q_j^n(0) = (u_j - a_L)e^{-r_j\tau_j/\beta}$ and τ_j denotes the end of firm j's (private) bank that solves (see (16))

$$\frac{(a_H - a_L)T}{u_j - a_L} = \tau_j - \frac{\beta}{r_j}\left(1 - e^{-r_j\tau_j/\beta}\right) \tag{18}$$

Note that $\tau_j(u_j)$ is a decreasing function of u_j,[17] so there may be high levels of u_j for which $\tau_j < T$ and for which the firm does not bank any permits.

Since

$$\dot{Q}^\circ(t) = \frac{r}{\beta}Q^p(t) + \frac{r_1}{\beta}q_1^n(t) + \frac{r_2}{\beta}q_2^n(t) \tag{19}$$

where Q^p is the total reduction from participating firms (superscript "p" stands for participation in the market), when both r_1 and r_2 are either greater or smaller than r, it is immediate that $Q^\circ(t)$ will always differ from $Q^*(t)$. When $r_1 < r < r_2$, $\dot{Q}^\circ(t)/Q^\circ(t) = r/\beta$

[17] Taking the total derivative of (18) with respect to u and rearranging yields (index j is omitted)

$$\frac{d\tau}{du} = \frac{-(a_H - a_L)T}{(u - a_L)^2(1 - e^{-r\tau/\beta})} < 0$$

only if

$$(r - r_1)q_1^n(t) = (r_2 - r)q_2^n(t) \tag{20}$$

holds for all t. Let suppose that (20) holds for t, then from (17), we have that in $t + \Delta$ (where Δ is very small) $q_1^n(t + \Delta)/q_1^n(t) = 1 + r_j\Delta/\beta$ and

$$(r - r_1)q_1^n(t + \Delta) \equiv (r - r_1)\left(1 + \frac{r_1\Delta}{\beta}\right)q_1^n(t) < (r_2 - r)\left(1 + \frac{r_2\Delta}{\beta}\right)q_2^n(t) \equiv (r_2 - r)q_2^n(t + \Delta) \tag{21}$$

Consequently, when $r_1 < r < r_2$, Q° will also differ from Q^*.[18] In sum, when all non-participating firms discount the future at some rate different than r, the evolution of Q° will differ from Q^* since the beginning of the banking program.

Consider now case (b). If $r_j = r$ for all j (where j still refers to a non-participating firm), then $\dot{Q}^\circ/Q^\circ = r/\beta$ for all t. Because we cannot a priori rule out $q_j^n(0) = q_j^*(0)$ for many combinations of u_j and α_j,[19] Q° and Q^* may in fact follow close paths during the early periods, at least, of the banking program. As the end of the efficient banking program τ^* is approached, Q° could still follow Q^* if and only if $\tau_j = \tau^*$ for all j. Because of heterogeneity in firms' unrestricted emissions, however, the latter is unlikely to happen. In fact, when $u_i \neq u_k$ for all $i = 1, ..., n$ and $i \neq k$ and τ_i is the end of firm i's (private) bank according to (18), the latter leads to $\tau_i \neq \tau_k$ and $\tau_i = \tau^*$ only for $u_i = \overline{u}$.

It is very unlikely, then, that the end of each non-participating firm's bank will ever coincide with the end of the efficient banking program. If $\tau_j > \tau^*$ for at least one non-participating firm, then $Q^\circ(\tau^*) < Q^*(\tau^*)$. If, on the other hand, $\tau_j < \tau^*$ for all j (i.e.,

[18] One might argue that a (sub-optimal) observed path can still follow the optimal path if we let βs vary across firms such that the rate r_j/β_j for each non-participating firm equals the aggregate rate "r/β". Because this aggregate rate varies with time when βs differ across firms, however, it will also differ from r_j/β_j.

[19] This combination derives from setting

$$\frac{q_j^*(0)}{q_j^n(0)} = \frac{U - A_L}{u_j - a_L}\left(\frac{\gamma}{\alpha_j}\right)^{1/\beta} = 1$$

$u_j > \overline{u}$ for all j), then the end of the banking period for the group of participating firms will be $\tau^p > \tau^*$ because average unrestricted emissions from the participating firms are now lower than $\overline{u}$. Consequently, we again have that $Q^\circ(\tau^*) < Q^*(\tau^*)$. The results of cases (a) and (b) can be summarized in the following proposition

Proposition 1 *The observed quantity path $Q^\circ(t)$ during a banking period will always differ from the optimal path $Q^*(t)$ if there exists at least one non-participating firm j for which either $r_j \neq r$ or $u_j \neq \overline{u}$.*

Because it is most unlikely that one or more non-participating firms will have the same level of unrestricted emissions $\overline{u}$,[20] Proposition 1 indicates that even if, on average, non-participating firms do have unrestricted emissions equal to $\overline{u}$, the effects of their non-participation on the evolution of $Q(t)$ do not cancel out. Thus, the evolution of $Q(t)$ provides sufficient information to judge the overall market performance. Furthermore, if we allow u_i to vary over time and β to vary across firms, the range of possible values that τ_j can take for each potential non-participating firm will expand, making τ_j and τ^* even more likely to differ.

A natural question that Proposition 1 raises for the case in which all non-participating firms discount the future at r is whether differences between $Q^\circ(t)$ and $Q^*(t)$ can be detected early in the banking program or only towards its end. This question is not irrelevant for a banking program that is expected to last many years, such as that for the SO_2 market.[21] In this regard, Appendix A establishes

Proposition 2 *If all non-participating firms discount the future at r but there exists at least one non-participating firm j for which $u_j \neq \overline{u}$, then $Q^\circ(0) > Q^*(0)$.*

Simply stated, Proposition 2 says that non-participation, if it exists at all, has an immediate effect on the evolution of $Q(t)$ and $B(t)$, and therefore we would not need to collect data for the entire banking period before concluding about overall market performance. Figure 1 presents the efficient quantity path $Q^*(t)$ (path A) and a hypothetical

[20] As we will see later, in the case of the SO_2 market, Ellerman et al. (2000) document significant heterogeneity in u.

[21] According to Ellerman et al. (2000) and Ellerman and Montero (2002) the SO_2 bank is not expected to end before 2008.

observed path $Q^{\circ}(t)$ with non-participants discounting at r (path B). The end of the bank along the efficient path is τ^*, and the end of the bank for participating firms is denoted by τ^p, which can also be greater than or equal to τ^*. Non-participating firms' individual banks end anywhere between τ_l and τ_h (note that τ_h can be greater than, equal to, or smaller than τ^*).[22] Full compliance or exhaustion also requires that between 0 and τ_h, the cumulative reduction along path A must equal the cumulative reduction along path B. During the first years and before τ_l, both Q^* and Q° will grow at the same rate r/β. As some non-participants exhaust their individual banks, Q° will start growing more slowly and will eventually cross Q^* before τ^*; otherwise the compliance condition will not be satisfied.

Before illustrating the effect of non-participation on the equilibrium path with some numerical exercises based on data taken from the SO_2 permits market, I will examine in the next section another type of market imperfection that can also prevent a permits market from delivering the least-cost solution. We have already seen that the exercise of market power cannot be detected in a "static" context unless detailed cost information is obtained. I will next explore whether market power can be detected in a dynamic context.

4 Banking with market power

Consider now a market with banking in which there is a dominant firm and a competitive fringe.[23] The dominant firm acts as a von Stackelberg leader, and all firms in the fringe have perfect foresight. Consequently, the dominant firm's decision problem is to choose the price path along with its reduction (or emission) path that maximizes the net present value of its profits, knowing that each firm in the competitive fringe will take such a price path as given and that neither its bank nor the fringe's bank can go negative.

Although this problem has been solved already for a typical exhaustible resource

[22] As explained in the Appendix A, if $\tau_h < \tau^*$, then $\tau^p > \tau^*$ and Figure 1 will still apply after relabeling τ^p for τ_h and vice versa.

[23] Based on the analysis of Lewis and Schmalensee (1980) for an oligopolistic market, considering two or more large firms and a competitive fringe should not qualitatively alter the main result of this section.

under different set of assumptions (Salant, 1976; Gilbert, 1978; Newbery, 1981), the proposed solutions do not immediately apply to a permits bank for several reasons. First, extraction costs for permits are zero. Second, storage costs for permits are zero so speculators (and firms in the fringe) will make sure that prices neither jump nor grow at a rate higher than r. This also enables the dominant firm to buy permits from the fringe and store them for future use at no cost other than the opportunity cost of selling them earlier. Third, in a permits market, the dominant producer can still exercise market power after its stock (i.e., bank) and that of the fringe have been exhausted. So, contrary to what would occur in a typical exhaustible resource market, the dominant firm may still use its strategic position of the end of the banking period to exercise some market power during the banking period even if it does not receive any permits from the stock $(A_H - A_L)T$, but only an allocation flow throughout. Fourth, because the demand for permits comes not from a third party (e.g., consumers) but internally from the fringe and the dominant producer, the dominant firm's decision problem is the choice of not only a permits sale/purchase path (or a price path supported by a sales path), but also an abatement (or demand) path.

Rather than attempt a complete characterization of equilibrium paths for any possible permits allocation and cost structure,[24] I shall describe the equilibrium path for what seems to be the most general case. Let f index the competitive fringe and m represent the dominant producer. Abatement costs are as before, so the fringe and leader's cost curves are $C_f(Q_f(t)) = \gamma_f[Q_f(t)]^{\beta+1}$ and $C_m(Q_m(t)) = \gamma_m[Q_m(t)]^{\beta+1}$, respectively. Total permits allocations are also as before, although it is useful to make an artificial distinction here between stock and flow allocations.[25] The total flow (or per period) allocation is A_L, beginning in $t = 0$, and the total stock allocation is $(A_H - A_L)T$. The fringe receives fractions θA_L of the flow allocation and $\mu(A_H - A_L)T$ of the stock allocation, so the dominant firm receives $(1 - \theta)A_L$ and $(1 - \mu)(A_H - A_L)T$, respectively. I also assume that θ and γ_m/γ_f are small enough that the dominant producer is a seller of permits at

[24] In Liski and Montero (2002) we consider other allocation, cost and commitment structures. For instance, we consider a dominant firm (possibly a broker) that holds a large part of the permits stock but does not pollute.

[25] The stock is the cumulative number of permits allocated above the long-term goal of A_L.

the end of the banking period.[26]

Let us first consider the case in which $\mu = 0$. Under this permits allocation, the fringe, on the one hand, does not build a bank on its own, but buys permits from the dominant producer since the first period. The dominant firm, on the other hand, finds it profitable to build and manage a permits bank. Formally, the dominant firm solves

$$\max \quad \int_0^\infty [P(t)X(t) - C_m(Q_m(t))]\mathrm{e}^{-rt}dt \tag{22}$$

$$\text{s.t.} \qquad P(t) = C_f'(Q_f(t)) \tag{23}$$

$$X(t) = U_f(t) - Q_f(t) - A_f(t) \tag{24}$$

$$\dot{B}_m(t) = A_m(t) - U_m(t) + Q_m(t) - X(t) \qquad [\lambda_m(t)] \tag{25}$$

$$B_m(t) \geq 0 \qquad [\phi_m(t)] \tag{26}$$

$$B_m(0) = 0 \tag{27}$$

where $X(t)$ is the number of permits sold by the dominant firm during period t,[27] $B_m(t)$ is the dominant firm's bank, and λ_m and ϕ_m are multiplier functions.

Since firms in the fringe are price takers, it is irrelevant whether the leader solves for $P(t)$ or $Q_f(t)$. Replacing (23) and (24) in the objective function to form the corresponding Hamiltonian $H(Q_f, Q_m)$, the necessary conditions for optimality include satisfaction of (23)–(27) and

$$\frac{\partial H}{\partial Q_f} = [C_f''(Q_f(t))X(t) - C_f'(Q_f(t))]\mathrm{e}^{-rt} + \lambda_m(t) = 0 \tag{28}$$

$$\frac{\partial H}{\partial Q_m} = -C_m'(Q_f(t))\mathrm{e}^{-rt} + \lambda_m(t) = 0 \tag{29}$$

$$\dot{\lambda}_m(t) = -\frac{\partial H}{\partial B_m} = -\,\phi_m(t),\ \phi_m \geq 0,\ \ \phi_m B_m = 0 \tag{30}$$

From (28) and (29) we obtain

$$[C_f'(Q_f(t)) - C_f''(Q_f(t))X(t) - C_m'(Q_m(t))]\mathrm{e}^{-rt} = 0 \tag{31}$$

[26] The same qualitative results apply if the dominant firm is a monopsonist at the end of the banking period (the end or "choke" price will be lower than the competitive price).

[27] If the dominant firm acts as a monopsonist, then $X(t) < 0$.

Eq. (31) shows that if the strategy of the dominant firm is optimal, the discounted value of marginal revenues, $C_f' - C_f'' X$,[28] minus marginal costs must be the same in all periods during which the dominant firm sells (i.e., marginal revenues net of marginal costs must rise at the rate of interest). Furthermore, since the dominant firm continues to enjoy market power after both its stock and the fringe's stock are consumed, marginal revenues must be equal to marginal costs in all periods. At the end of the banking period τ^m, the "choke" price P^m that prevails does not depend on the allocations before T, and can be readily estimated by solving (31) subject to (24) and $Q_m(\tau^m) = U(\tau^m) - A_L - Q_f(\tau^m)$. Since the dominant firm is assumed to be a seller of permits in the long run we have $P^m > P^*(\tau^*)$.

A characterization of the price path during the banking period can be obtained from (28). Taking the derivative with respect to time, letting $\dot{\lambda}_m = 0$, and rearranging yields

$$\dot{P}(t) = rP(t) + \dot{C}_f''(Q_f(t))X(t) - rC_f''(Q_f(t))X(t) - C_f''\dot{Q}_f(t) \tag{32}$$

Because there are no storage costs, we already know that arbitrage prevents prices from increasing at any rate higher than the discount rate r; hence, $\dot{Q}_f/Q_f$ cannot be higher than r/β. Using this and $C_f(Q_f(t)) = \gamma_f[Q_f(t)]^{\beta+1}$, it is not difficult to show that

$$\frac{\dot{C}_f''}{C_f''} = (\beta - 1)\frac{\dot{Q}_f}{Q_f} \leq r$$

which, in turn, implies that $\dot{P}/P < r$.[29] Consistent with Salant (1976) and Newbery (1981), when the fringe has no stock left, it is optimal for the dominant producer to let prices rise at a rate strictly lower than the discount rate.

The quantity path, $Q(t) = Q_f(t) + Q_m(t)$, can be derived from the price path, eq. (31) and the exhaustion or full compliance conditions. From the price path and (31), we know that $\dot{Q}_f/Q_f < r/\beta$ and $\dot{Q}_m/Q_m = r/\beta$, respectively. This implies that in the presence of market power, the observed quantity path $Q^\circ(t)$ rises at a lower rate than does

[28] Note that since $C_f''(Q_f(t)) = \partial P(Q_f(t))/\partial Q_f(t)$, marginal revenues can be expressed as $P(t) - P'(X(t))X(t)$.

[29] Note that if marginal cost curves are linear, i.e., $\beta = 1$, $\dot{Q}_f/Q_f = \dot{P}/P$ and $\dot{P}/P = r/2$.

the efficient quantity path $Q^*(t)$. This result, together with the exhaustion conditions, indicates that $Q^\circ(t)$ must start above $Q^*(t)$ and cross it from above at some later point to finally converge to $U(\tau^m) - A_L$ at $\tau^m > \tau^*$. Although driven by different reasons, the effect of market power on $Q(t)$ is qualitatively similar to the effect of non-participation on $Q(t)$ that is depicted in Figure 1. Market power unambiguously prolongs the length of the banking period and increases the total cost of compliance.[30]

For this particular allocation of permits in which the fringe builds no stock, it is possible to detect the presence of market power by contrasting either $Q^\circ(t)$ with $Q^*(t)$ or $\dot{P}^\circ/P^\circ$ with r (recall that absolute price levels do not say much because of limited cost information). However, the latter becomes unfeasible when the fringe holds a bank. During the period in which the fringe's bank is positive, prices must rise at the interest rate; otherwise the fringe would not hold any permits.

Let us now consider the more general case in which the fringe holds a permits bank for some period of time. To make the case clear enough, let us assume that $\mu = 1$, so the dominant firm receives no stock. One can think of different candidates for the Stackelberg-rational expectations equilibrium. For example, the dominant firm could propose a price path growing at a lower rate that would induce firms in the fringe to sell all their stock as early as the first period. In the absence of binding contracts, however, this solution is time inconsistent, because as soon as the fringe's stock is exhausted, the dominant firm will find it profitable to revise its initial price path proposal and raise prices accordingly. Firms in the fringe will anticipate the price jump and hence hold onto their permits rather than sell them in the first place.

Since the dominant firm receives no stock, another candidate is one in which the dominant firm builds no bank and the fringe's bank expires at the choke price P^m. If this were indeed the solution, market power could not be detected from either price or quantity data. Price would rise at the rate of interest during the entire banking period, and from the exhaustion conditions it is clear that the aggregate quantity path $Q(t) = Q_f(t) + Q_m(t)$ would coincide with the competitive path $Q^*(t)$. This solution

[30] This is in contrast with Stiglitz (1976) and Weinstein and Zeckhauser (1979) who show that for a typical exhaustible resource the effect of market power on the direction and magnitude of the departure from optimality cannot be predicted in general.

cannot be an equilibrium either, because the dominant firm sells permits before the end of the banking period. Since the dominant firm has enough flexibility to support this price path through different sales paths (all yielding the same discounted sum of profits of the fringe and the leader), it can certainly choose to accelerate the exhaustion of the fringe's bank by holding onto its permits and selling them only after the fringe bank has been exhausted at $\tau^f < \tau^*$. At τ^f, however, the dominant firm would find its original proposal no longer optimal and would let prices rise (after a possible instantaneous jump) at a rate strictly lower than r until they reach P^m at $\tau^m > \tau^* > \tau^f$.

Consequently, the equilibrium path must necessarily have the dominant firm conserving enough permits to keep a stock that will consume and sell after all firms in the fringe have exhausted theirs, regardless of how much it received of the stock $(A_H - A_L)T$. Before providing the complete solution, the latter equilibrium condition gives us sufficient information to depict typical equilibrium price and quantity paths, $P^\circ(t)$ and $Q^\circ(t)$, respectively. As shown in Figures 2 and 3, there will be three distinctive phases. During phase A, $P^\circ(t)$ rises at the interest rate r, and $Q_f(t)$ and $Q_m(t)$ rise at r/β, as in the competitive case. While the fringe consumes its stock and the dominant firms builds its own, it is not obvious whether the dominant firm participates in the market during this phase (more on this below). At τ^f, the fringe's bank is exhausted but the dominant firm's bank is positive. Phases B and C are as before. In phase B, $P^\circ(t)$ rises at a rate strictly lower than r, and $Q^\circ(t) = Q_f(t) + Q_m(t)$ grows at a rate strictly lower than r/β since $Q_f(t)$ follows the price path. Furthermore, from the full compliance (or exhaustion) condition, the observed path $Q^\circ(t)$ crosses the competitive path $Q^*(t)$ sometime during this phase. At τ^m, the leader's bank is exhausted, after which prices remain constant at $P^m > P^*$.

Because quantity data allow us to detect market power even when prices rise at rate r during phase A, the results of this section can be summarized in the following proposition:

Proposition 3 *In the presence of market power, $Q^\circ(t) \neq Q^*(t)$, regardless of the allocation of the permits stock, i.e., $(A_H - A_L)T$. More specifically, $Q^\circ(0) > Q^*(0)$ and $\tau^m > \tau^*$.*

Although Proposition 3 establishes that market power will immediately affect the

quantity path (which answers one central question of this paper), it does not say much about the difference in magnitude between $Q^{\circ}(t)$ and $Q^{*}(t)$ for a given permits allocation (θ and μ) and cost structure (γ_f and γ_m). For that we must derive the complete equilibrium solution. In particular, we need to determine τ^f and τ^m.

The solution must not only be time consistent and exhibit the market power of the dominant firm after the fringe's bank has expired, but one can argue that it should also make some use of the ability of the dominant firm to alter the stock of the fringe during the competitive phase (phase A) by either selling or buying permits. In the absence of binding contracts, however, the latter possibility will be time inconsistent in the sense that the dominant firm would continuously like to revise its original price path after each transaction.[31] To overcome these objections and still allow the dominant firm to be more active during the competitive phase, Newbery (1981) argued that the Nash-Cournot equilibrium appears to be the best approximation to the rational expectations Stackelberg equilibrium.[32] In our context, however, such an approximation looks less attractive to the leader, since in a permits market where there is no third-party demand, the Nash-Cournot equilibrium coincides with the Nash bargaining solution (Spulber, 1989) in which $P = C_f'(Q_f) = C_m'(Q_m)$. Hence, the more reasonable solution is for the dominant firm to refrain from any permits transaction during the competitive phase and only start selling permits at τ^f (the formal derivation of such an equilibrium solution can be found in Appendix B).

5 An application

The U.S. SO_2 trading program is a natural candidate for demonstrating the application of the model. Hence, the purpose of this section is not to present a formal efficiency test

[31] Since the dominant firm's optimal sale or purchase is a function of the fringe' stock, the ex-ante (i.e., before the transaction) optimal solution differs from the ex-post optimal solution. The rational expectations Stackelberg equilibrium derived by Gilbert (1978) in his example does not have this time inconsistency problem because he uses a constant demand elasticity (besides equal discount rates and zero extraction costs), which is not our case. See Liski and Montero (2002) for more on this time inconsistency issue.

[32] The Nash-Cournot equilibrium is also used by Salant (1976).

for the actual evolution of the SO_2 bank,[33] but to illustrate the effect of different forms of non-participation (or inefficient participation) on the evolution of prices and quantities and see whether it is possible to detect non-participation with some degree of confidence. Furthermore, because market power does not seem to be an issue in the SO_2 program, I focus exclusively on the effects of non-participation.

For the application, I use data from the 263 electric utility power plants affected in both Phase I, which lasted for $T = 5$ years, and Phase II of the program, which is ongoing. The total number of permits (or allowances) allocated to all of these plants each year during Phase I was $A_H = 6.31$ million and during Phase 2 was $A_L = 2.37$ million. Each permit gives its holder the right to emit one ton of SO_2 in a particular year. To allow for the possibility that a plant owner may partially participate in the market (i.e., engage in some trading activity, but not enough to equate marginal costs and prices), I divide each of the 263 plants into smaller production units of roughly 100 MW each,[34] resulting in a total of $n = 881$ units. I then treat each of these production units as an independent firm that either fully participates in the market or does not participate at all, as in the model.

Statistics for the 881 units are summarized in Table 1. Permits allocations for each unit (a_H^i and a_L^i) are obtained by dividing the allocation of the original plant to which the unit belongs by the number of production units in that plant. Individual unrestricted or counterfactual emissions (u_i) are obtained in a similar way, and are approximately equal to emissions at the time the SO_2 program was signed into law in 1990 multiplied by a 10-year growth factor of 6.5%, based on EPA's emissions forecast at that time.[35] Counterfactual emissions total $U = 9.14$ million tons. Cost parameters for each unit (α_i) are randomly assigned from a uniform distribution over the interval $[0.002; 0.0002]$. To be consistent with previous estimates, these numbers were chosen to produce an initial equilibrium price of about \$260 and long-run (i.e., after the banking period) cost savings

[33] A formal test would require an empirical estimate of several parameters including the discount rate. For more see Ellerman and Montero (2002).

[34] For example, a plant 430-MW plant is converted into 4 smaller units while a 670-MW plant is converted into 7 smaller units. Plant size ranges from 100 MW to 1500 MW.

[35] For 21 units, I increased counterfactual emissions a bit further just to avoid corner solutions. On aggregate, this represents a less than 1% increase of counterfactual emissions.

from trading on the order of 45%. In addition, I use convex marginal cost curves with $\beta = 1.5$ (in order to avoid corner solutions) and a discount rate of $r = 6\%$. To keep things simple, I assume that a firm j that does not participate is one of three possible types, depending on its discount rate: (1) $r_j = r$, (2) $r_j = r/2$ and (3) $r_j \gg r$ (i.e., no banking).

Simulation results for relevant variables and different levels of market participation are in Table 2. Since each market simulation randomly assigns to each unit a cost parameter α_i, a participation status, and a non-participation type, the results presented are averages over several simulation runs. As a benchmark, the first row shows the command-and-control (CAC) solution in which firms are prevented from engaging in both spatial and intertemporal trading, hence; $Q(0)$ and $Q(\tau)$ are the total reductions in each year during Phase I and Phase II, respectively, and $C(0)$ and $C(\tau)$ are the corresponding costs. The second row presents the market efficient solution, with a banking period (τ) of about 13 years and a bank at the end of Phase I, $B^*(T)$, of 8.04 million permits.

Without any prior regarding the proportion of types among non-participants, the third row shows the effects of a 25% non-participation rate, assuming that non-participation types are in equal proportions. While by the end of Phase I the actual bank $B^\circ(T)$ is only 5% smaller than $B^*(T)$, by the tenth year the actual bank $B^\circ(10)$ is 18% larger than the efficient bank $B^*(10)$; this result is comparable to the 18% long-run efficiency losses (i.e., higher costs $C(\tau)$) from less than optimal market participation. As shown in rows 4 and 5, these differences between the actual bank path and the efficient path increase steadily as the participation rate falls.

The next seven rows (from 6 to 12) present results for a 75% participation rate and different combinations of non-participation types. Differences in both the levels and the rates of change between the actual and the efficient bank path are always important, particularly as the bank is withdrawn. For instance, row 9 considers a proportion of non-participation types (3/4 of type 2 and 1/4 of type 3) such that $B^\circ(T)$ is almost equal to $B^*(T)$, but because the rate at which the efficient bank is withdrawn is higher than the rate at which the actual bank is withdrawn, $B^\circ(10)$ is considerably larger than $B^*(10)$. Interestingly, when all non-participants discount the future at the market rate r (i.e.,

type 1), as in row 10, the differences between the efficient and the actual path reduce significantly, but not so fully as to prevent the detection of some market inefficiencies ($B^{\circ}(10)$ is still 13% larger than $B^{*}(10)$).

Because actual and efficient paths differ at various point in time, precise knowledge of total unrestricted (or counterfactual) emissions, as we have assumed so far, is not crucial for the detection of non-participation, if it exists. Row 13 shows the efficient solution for a market in which total unrestricted emissions are assumed to be 10% higher than before (i.e., 10.05 million tons).[36] Thus, if we perform an efficiency test assuming incorrectly that total unrestricted emissions are 10.05 instead of 9.14, the effect on $B(t)$ of a 60% participation rate (with equal proportion of non-participation types) may not be detected easily at T, as shown in row 14. However, since an actual bank with partial participation and the efficient bank evolve at different rates, their paths will inevitably differ both before and after T.[37] This applies to different participation rates and to different proportions of non-participation types, as well.

6 Final Remarks

I have investigated the effects of less than optimal market participation and of the exercise of market power on the equilibrium path of a permits market with banking. During the period in which firms bank and withdraw permits from the bank (i.e., the banking period), the efficient price path follows Hotelling's rule, rising at the interest rate; because of imperfect cost information, however, the actual price path does not provide enough information to detect either non-participation or market power (at least during the competitive phase). The efficient permits bank path, on the other hand, is unique and can be readily contrasted with the evolution of the actual bank. In the case of non-participation, this is possible because of heterogeneity across firms (particularly in term of their counterfactual emissions). In the case of market power, this is possible because the dominant firm always conserves a stock of permits after all firms in the fringe have

[36] To facilitate the numerical solution of the model without affecting the aggregate quantity results, I also assume that all individual unrestricted emissions increase by 10%.

[37] Note that $B(0) = A_H - U + Q(0)$.

exhausted theirs. After that, the price path rises at a rate strictly lower than the interest rate until it reaches the static monopoly level.

In an effort to contribute to the aforementioned debate about the performance of the permits market of the U.S. Acid Rain Program, I then applied the theoretical model to data obtained from the SO_2 permits program. Numerical exercises indicate that the levels of inefficiency (i.e., higher compliance costs) suggested by Carlson et al. (2000) can be supported only by a significant degree of non-participation: about 50%, if we believe that non-participants use discount rates that are not necessarily equal to the market discount rate, as shown in row 4 of Table 2. Furthermore, for this degree of non-participation, the evolution of the actual bank would differ noticeably from the evolution of the optimal bank; it would be 11% smaller by $t = 5$ (end of Phase 1) and 36% larger by $t = 10$ (see row 4). This departure from optimal banking is in sharp contrast with the empirical analysis of Ellerman and Montero (2002), who found that the aggregate evolution of the SO_2 bank for the period 1995-2001 closely follows the evolution of an optimal bank.

References

[1] Carlson, C., D. Burtraw, M. Cropper, and K. Palmer (2000), Sulfur dioxide control by electric utilities: What are the gains from trade? *Journal of Political Economy* 108, 1292-1326.

[2] Ellerman, A.D, P. Joskow, R. Schamlensee, J.-P. Montero, and E.M. Bailey (2000), *Markets for Clean Air: The US Acid Rain Program*, Cambridge University Press, New York.

[3] Ellerman, A.D. and J.-P. Montero (2002), The temporal efficiency of SO_2 emissions trading, working paper, MIT-CEEPR.

[4] Gilbert, R.J. (1978), Dominant firm pricing policy in a market for an exhaustible resource, *Bell Journal of Economics* 9, 385-95.

[5] Hahn, R. (1984), Market power and transferable property rights, *Quarterly Journal of Economics* 99, 753-765.

[6] Hotelling, H. (1931), The economics of exhaustible resources, *Journal of Political Economy* 39, 137-175.

[7] Joskow, P., R. Schmalensee, and E. Bailey (1998), The market for sulfur dioxide emissions, *American Economic Review* 88, 669-685.

[8] Kamien, M.I. and N.L. Schwartz (1991). *Dynamic Optimization: The Calculus of Variation and Optimal Control in Economics and Management*, North-Holland, New York.

[9] Lewis, T. and R. Schmalensee (1980), On oligopolistic markets for nonrenewable resources, *Quarterly Journal of Economics* 95, 475-491.

[10] Liski, M. and J.-P. Montero (2002), Market power in pollution permit banking, mimeo, Helsinki School of Economics, Helsinki.

[11] Montero, J.-P. (1998), Marketable pollution permits with uncertainty and transaction costs, *Resource and Energy Economics* 20, 27-50.

[12] Montgomery, W.D. (1972), Markets in licenses and efficient pollution control programs, *Journal of Economic Theory* 5, 395-418.

[13] Newbery, D.M. (1981), Oil prices, cartels and the problem of dynamic inconsistency, *Economic Journal* 91, 617-646.

[14] Rubin, J.D. (1996), A model of intertemporal emission trading, banking, and borrowing, *Journal of Environmental Economics and Management* 31, 269-286.

[15] Salant, S.W. (1976), Exhaustible resources and industrial structure: A Nash-Cournot approach to the world oil market, *Journal of Political Economy* 84, 1079-1093.

[16] Schennach, S.M. (2000), The economics of pollution permit banking in the context of Title IV of the 1990 Clean Air Act Amendments, *Journal of Environmental Economics and Management* 40, 189-210.

[17] Spulber, D. (1989), *Regulation and Markets*, MIT Press, Cambridge, Massachusetts.

[18] Stavins, R. (1995), Transaction costs and tradeable permits, *Journal of Environmental Economics and Management* 29, 133-148.

[19] Stiglitz, J.E. (1976), Monopoly and the rate of extraction of exhaustible resources, *American Economic Review* 66, 655-661.

[20] Weinstein, M.C. and R.J. Zeckhauser (1975), The optimal consumption of depletable natural resources, *Quarterly Journal of Economics* 89, 371-392.

[21] Weitzman, M. (1974), Prices vs. quantities, *Review of Economic Studies* 41, 477-491.

A Proof of Proposition 3

I provide a general proof by considering two special cases: (1) two non-participating firms 1 and 2, for which $u_1 < \overline{u} < u_2$; and (2) one non-participating firm j, for which $u_j \neq \overline{u}$.

Case (1): There are only two non-participating firms 1 and 2 for which $u_1 < \overline{u} < u_2$. From (16) and (18) we have that $\tau_1 > \tau^* > \tau_2$. In addition, if we assume, for notational simplicity, that $u_1 < (U - u_1 - u_2)/(n-2) < u_2$,[38] we also have that $\tau_1 > \tau^p > \tau_2$, where τ^p is the end of the bank for the group of participating firms.

Full compliance requires

$$\int_0^{\tau_1} Q^*(t)dt = \int_0^{\tau_1} Q^\circ(t)dt \tag{33}$$

Replacing $Q^\circ(t)$ by $q_1^n(t) + q_2^n(t) + Q^p(t)$, where Q^p is the total reduction from the group of participating firms, (33) becomes (note that, to save on notation, I have made the time index a subscript for $t = 0$)

$$\int_0^{\tau^*} Q_0^* \mathrm{e}^{rt/\beta}dt + (U - A_L)(\tau_1 - \tau^*) = \int_0^{\tau_1} q_{10}^n \mathrm{e}^{rt/\beta}dt + \int_0^{\tau_2} q_{20}^n \mathrm{e}^{rt/\beta}dt + (u_2 - a_L)(\tau_1 - \tau_2) + \int_0^{\tau^p} Q_0^p \mathrm{e}^{rt/\beta}dt + (U - u_1 - u_2 - A_L + 2a_L)(\tau_1 - \tau^p) \tag{34}$$

[38] This assumption is equivalent to saying that $(u_1 + u_2)/2$ is not too different from $\overline{u}$.

Developing (34), using each bank's terminal condition (see (14)), and rearranging terms leads to

$$f(u_1,u_2) \equiv \frac{\beta}{r}[q_{10}^n(u_1)+q_{20}^n(u_2)+Q_0^p(u_1,u_2)-Q_0^*] = \\ (U-u_1-u_2-A_L+2a_L)(\tau_1-\tau^p)+(u_2-a_L)(\tau_1-\tau_2)-(U-A_L)(\tau_1-\tau^*) \quad (35)$$

Thus, the proof for case (1) would be complete if we can demonstrate that $f(u_1,u_2)>0$. Although the first two terms on the right-hand side (RHS) of (35) are positive, the third term is negative, so the sign of $f(u_1,u_2)$ remains ambiguous.

I proceed the demonstration with a comparative static analysis by letting $u_1=\overline{u}-\Delta$ and $u_2=\overline{u}+\Delta$, with Δ sufficiently small to use a second-order Taylor's approximation for $f(u_1,u_2)$ around $f(\overline{u},\overline{u})$, as follows

$$f(u_1,u_2) \approx f(\overline{u},\overline{u})+f_1'(\overline{u},\overline{u})(-\Delta)+f_2'(\overline{u},\overline{u})\Delta \\ +\frac{1}{2}[f_{11}''(\overline{u},\overline{u})\Delta^2+2f_{12}''(\overline{u},\overline{u})(-\Delta^2)+f_{22}''(\overline{u},\overline{u})\Delta^2] \quad (36)$$

where subscripts 1 and 2 in f denote (partial) derivatives with respect to u_1 and u_2, respectively. Since Q_0^p is unaffected by Δ because $(u_1+u_2)/2=\overline{u}$, then $f_1'(\overline{u},\overline{u})=f_2'(\overline{u},\overline{u})$. Thus, substituting $f(\overline{u},\overline{u})=0$ and $f_{12}''(\overline{u},\overline{u})=0$ into (36) we obtain

$$f(u_1,u_2)=\frac{\Delta^2}{2}[f_{11}''(\overline{u},\overline{u})+f_{22}''(\overline{u},\overline{u})]=\frac{d^2\left(q_{k0}^n(u_k=\overline{u})\right)}{du_k^2}\Delta^2 \quad (37)$$

where k is either 1 or 2.

Thus, we must now demonstrate that q_{k0}^n is convex in u_k. Since $q_{k0}^n=(u_k-a_L)\mathrm{e}^{-r\tau_k/\beta}$, plugging the latter into (18) we obtain (hereafter I drop the indices k and n)

$$(a_H-a_L)T=(u-a_L)\left(\tau-\frac{\beta}{r}\right)+\frac{\beta}{r}q_0$$

which, after total differentiating by u, yields

$$0=\tau-\frac{\beta}{r}+(u-a_L)\frac{d\tau}{du}+\frac{\beta}{r}\frac{dq_0}{du} \quad (38)$$

Obtaining $d\tau/du$ by taking the total derivative of (18) with respect to u, substituting it into (38), and rearranging terms leads to

$$\frac{dq_0}{du} = \frac{r\tau/\beta}{\mathrm{e}^{r\tau/\beta} - 1} > 0$$

and

$$\frac{d^2q_0}{du^2} = \frac{r}{\beta}\frac{\left(\mathrm{e}^{r\tau/\beta} - 1 - \frac{r}{\beta}\tau\mathrm{e}^{r\tau/\beta}\right)}{\left(\mathrm{e}^{r\tau/\beta} - 1\right)^2}\frac{d\tau}{du}$$

Since $d\tau/du < 0$, it remains to be demonstrated that the expression in parentheses in the numerator is negative. Multiplying (18) by $\mathrm{e}^{r\tau/\beta}$ and rearranging, we obtain

$$e^{r\tau/\beta} - 1 - \frac{r}{\beta}\tau\mathrm{e}^{r\tau/\beta} = -\frac{r(a_H - a_L)}{\beta(u - a_L)}\mathrm{e}^{r\tau/\beta} < 0$$

which finishes the proof for Case (1).

Case (2): There is only one non-participating firm j for which $u_j = \overline{u} + \Delta$ (we shall see that the same result is obtained for $u_j = \overline{u} - \Delta$). Since $\tau_j < \tau^* < \tau^p$, full compliance requires

$$\int_0^{\tau^p} Q^*(t)dt = \int_0^{\tau^p} Q^\circ(t)dt \tag{39}$$

Replacing $Q^\circ(t)$ by $q_j^n(t) + Q^p(t)$ in (39) and proceeding as before we obtain

$$f(u_j) \equiv \frac{\beta}{r}[q_{j0}^n(u_j) + Q_0^p(u_j) - Q_0^*] = (u_j - a_L)(\tau^p - \tau_j) - (U - A_L)(\tau^p - \tau^*) \tag{40}$$

Since the sign of $f(u_j)$ remains ambiguous from (40), I again let Δ be sufficiently small in order to use a second-order Taylor's approximation for $f(u_j)$ around $f(\overline{u})$, as follows

$$f(u_j) = f(\overline{u}) + f'(\overline{u})\Delta + \frac{1}{2}f''(\overline{u})\Delta^2 \tag{41}$$

To compute f' and f'', it is useful to write q^n_{j0} and Q^p_0 as functions of u_j (or Δ), as follows

$$q^n_{j0}(u_j) \equiv q^n_{j0}(x) = (\overline{u} + x - a_L)\mathrm{e}^{-r\tau(x)/\beta} \tag{42}$$

$$Q^p_0(u_j) \equiv Q^p_0(y) = (n-1)(\overline{u} + y - a_L)\mathrm{e}^{-r\tau(y)/\beta} \tag{43}$$

where $x = \Delta$, $y = -\Delta/(n-1)$, and $\tau(z = x, y)$ is obtained from

$$\frac{(a_H - a_L)T}{\overline{u} + z - a_L} = \tau(z) - \frac{\beta}{r}\left(1 - \mathrm{e}^{-r\tau(z)/\beta}\right)$$

Since $dq^n_{j0}(u_j)/du_j = dq^n_{j0}(x)/dx$ by construction ($u_j = \overline{u} + x$), we have that

$$\frac{dQ^p_0(u_j)}{du_j} = \frac{dQ^p_0(y)}{dy}\frac{dy}{du_j} \tag{44}$$

Replacing $dy/du_j = -1/(N-1)$ and using the similarity between (42) and (43), we get

$$\frac{dQ^p_0(u_j)}{du_j} = (N-1)\frac{dq^n_{j0}(x)}{dx}\frac{-1}{N-1} = -\frac{dq^n_{j0}(u_j)}{du_j} \tag{45}$$

In addition, taking the derivative of (44) with respect to u_j yields

$$\begin{aligned}\frac{d^2Q^p_0(u_j)}{du_j^2} = \frac{d^2Q^p_0(y)}{dy^2}\left(\frac{dy}{du_j}\right)^2 + \frac{dQ^p_0(y)}{dy}\frac{d^2y}{du_j^2} \\ = (N-1)\frac{d^2q^n_{j0}(x)}{dx^2}\left(\frac{1}{N-1}\right)^2 = \frac{1}{N-1}\frac{d^2q^n_{j0}(u_j)}{du_j^2}\end{aligned} \tag{46}$$

Plugging $dq^n_{j0}(u_j)/du_j$, $d^2q^n_{j0}(u_j)/du_j^2$, (45), and (46) into (41) yields

$$f(u_j) = \frac{1}{2}f''(\overline{u})\Delta^2 = \frac{N\Delta^2}{2(N-1)}\frac{d^2q^n_{j0}(u_j = \overline{u})}{du_j^2} \tag{47}$$

Since q^n_{j0} is convex in u_j, as demonstrated in Case (1), $f(u_j) > 0$ when there is only one non-participating firm j for which $u_j = \overline{u} + \Delta$. Following the same procedure, it is immediate that (47) also results when $u_j = \overline{u} - \Delta$.

B The dominant firm's solution

The solution of the dominant firm is found by first imposing a continuous price path that ends at P^m, the monopoly price that prevails when there is no bank left and which can be easily obtained from (31). From 0 to τ^f (the time at which the fringe's bank is exhausted), $\dot{P}(t)/P(t) = r$, and from τ^f to τ^m (the time at which the dominant firm's bank is exhausted), $\dot{P}(t)/P(t) < r$, according to (32). At τ^m and afterward, $P(t) = P^m > P^*(\tau^m)$.

The rest of the solution (i.e., τ^m and τ^f) is found by simultaneously solving the two "exhaustion" conditions: the fringe's bank expires at τ^f and the dominant firm's bank expires at $\tau^m > \tau^f$. Since the dominant firm does not trade between 0 and τ^f, these two conditions can be written as

$$\int_0^{\tau^f} Q_f(t)dt = (U_f - \theta A_L)\tau^f - \mu(A_H - A_L)T \tag{48}$$

$$\int_0^{\tau^m} (Q_f(t) + Q_m(t))dt = (U - A_L)\tau^m - (A_H - A_L)T \tag{49}$$

where $U_f = U - U_m$ and θ and μ are, respectively, the proportion of flow and stock permits allocated to the fringe.

The fringe's abatement path $Q_f(t)$ follows the price path according to $C_f'(Q_f(t)) = P(t)$. The dominant firm's abatement path $Q_m(t)$, on the other hand, must minimize the present value of the dominant firm's compliance costs during the banking period; hence, it must grow at r/β until it reaches its long-term level Q_m^m at τ^m (this value can also be obtained from (31)). Substituting these abatement paths and the price path into (48) and (49), τ^f and τ^m are finally found.

TABLE 1. Summary statistics (in thousands)

Variable	# units	Mean	St. dev.	Min	Max	Total
a_H	881	7.17	2.05	0.42	12.49	6,314
a_L	881	2.69	0.64	0.14	4.60	2,372
u	881	10.37	3.85	0.40	22.63	9,135

TABLE 2. The effect of non-participation (Q, B, and C, in millions)

	Particip. Rate	Prop. Types (1);(2);(3)*	τ	Q(0)	Q(τ)	B(T)	B(10)	B(τ)	P(0)	C(0)	C(τ)
1	CAC	n.a.	n.a.	2.88	6.76	n.a.	n.a.	n.a.	n.a.	580	2817
2	100%	n.a.	13.13	4.00	6.76	8.04	1.27	0	259	414	1539
3	75%	1/3;1/3;1/3	13.11	3.97	6.69	7.64	1.50	0.22	258	535	1813
4	50%	1/3;1/3;1/3	13.16	3.94	6.62	7.19	1.73	0.44	261	656	2087
5	<1%	1/3;1/3;1/3	13.23	3.89	6.49	6.42	2.18	0.88	247	901	2640
6	75%	1/2;1/2;0	13.12	4.10	6.65	8.44	1.78	0.34	259	577	1788
7	75%	1/2;0;1/2	13.11	3.86	6.73	7.05	1.19	0.06	259	496	1852
8	75%	0;1/2;1/2	13.12	3.96	6.69	7.41	1.53	0.27	260	531	1785
9	75%	0;3/4;1/4	13.14	4.07	6.65	8.08	1.82	0.40	259	583	1783
10	75%	1; 0; 0	13.11	4.01	6.70	8.11	1.43	0.12	260	527	1824
11	75%	0; 1; 0	13.14	4.18	6.61	8.73	2.09	0.53	259	617	1746
12	75%	0; 0; 1	13.14	3.72	6.76	6.00	0.95	0	256	615	1986
13	100%	n.a.	12.25	4.70	7.68	7.35	0.76	0	328	617	2103
14	60%	1/3; 1/3; 1/3	13.16	3.96	6.64	7.41	1.65	0.35	256	615	1986

*Non-participation types are: (1) r = 6%, (2) r = 3%, and (3) r $\gg$ 6%. Counterfactual in row 13 is 10% higher

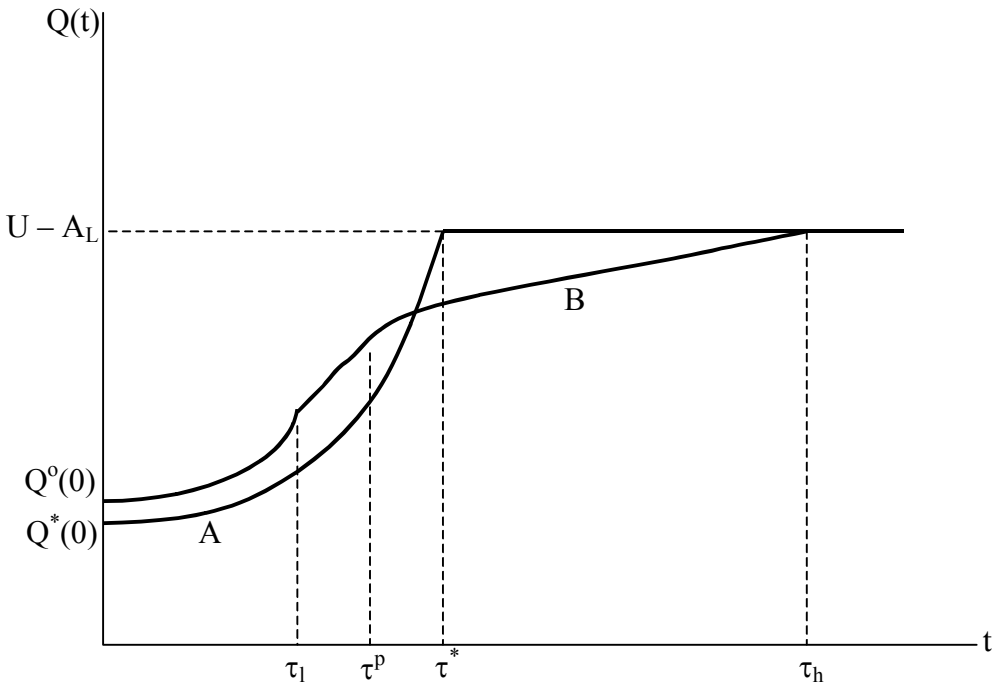

Figure 1: Effect of non-participation on the abatement path, $Q^{\circ}(t)$.

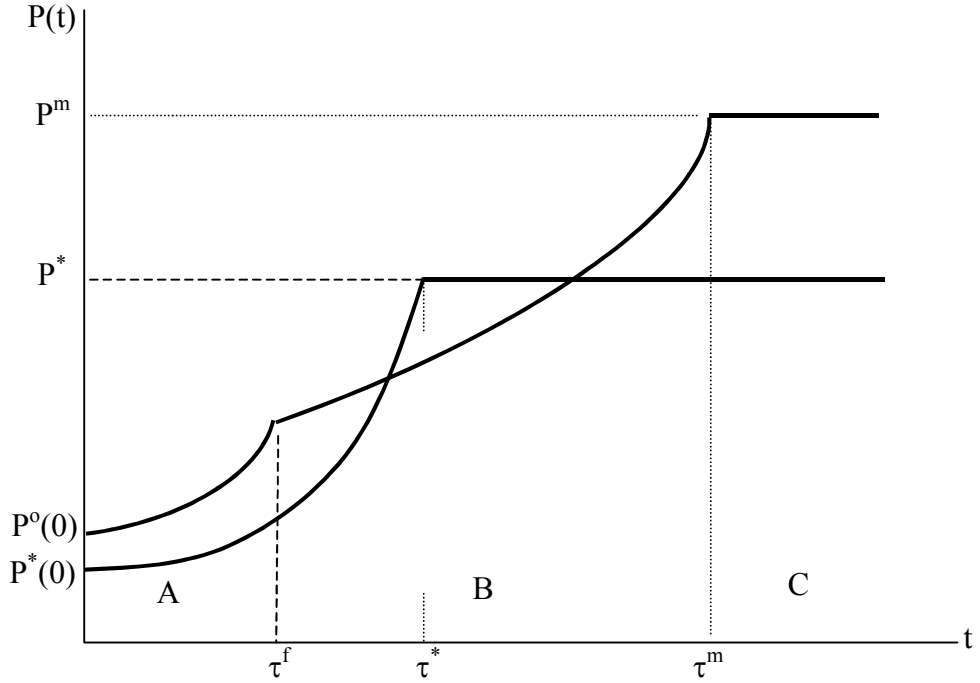

Figure 2: Effect of market power on the price path, $P^{\circ}(t)$.

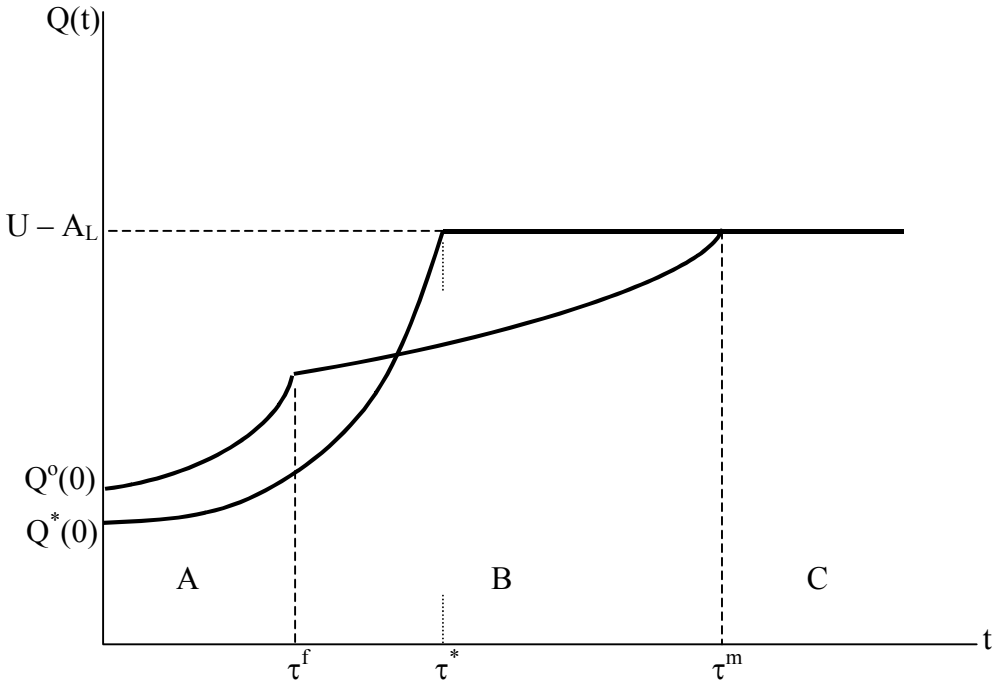

Figure 3: Effect of market power on the abatement path, $Q^{\circ}(t)$.