Arbitrage in Energy Markets: Competing in the Incumbent's Shadow

Gerd Küpper; Bert Willems*†

November 2007

Abstract

This paper studies the welfare implications of using market mechanisms to allocate transmission capacity in recently liberalized electricity markets. It questions whether access to this essential facility should be traded on a market, or whether the incumbent should retain exclusive usage rights. We show that granting exclusive use to the incumbent might be optimal, if the capacity of the essential facility is small and the incumbent can reduce production costs by taking advantage of interregional production-cost differences. This result counters the intuition that arbitrage will improve the social surplus when there is no output contraction. The reason is that when competition is imperfect, arbitrage might reduce production efficiency. We advise policymakers to introduce market mechanisms for the allocation of transmission capacity only if sufficient investment in the network is ensured or if the market power of the incumbent is broken in at least one of the markets in which it is active.
JEL D40, L10, L50,Q48
Keywords: Arbitrage, electricity sector, price discrimination

* *K.U.Leuven, C.E.S. and Energy Institute. Gerd.Kupper@econ.kuleuven.be*
† *Tilburg University, TILEC and CentER; K.U.Leuven, C.E.S. and Energy Institute. B.R.R.Willems@uvt.nl*

1 Introduction

The electricity sector has been subject to major structural changes during the last decade. Liberalization policies all over the world have led to a separation of formerly vertically integrated monopolies into three parts: production, retail and network services. Competition has been introduced at the production- and retail levels, although most markets remain highly concentrated and incumbent firms often continue to be dominant. Competition at the level of network services is not feasible, as it is inefficient to build and operate multiple parallel networks. These network assets are therefore essential facilities whose efficient allocation is crucial for a well-functioning upstream (production) and downstream (retail) market. Regulation generally requires that access to the essential facility be organized in such a way that it is non-discriminatory and market-conform. This implies that price arbitrage becomes possible, and that it is harder for the incumbent generation firm to price discriminate. This paper studies the welfare effect of a change towards a more market-conform allocation of essential facilities in the electricity sector, while at the same time keeping market structure (ownership) constant.

The classical industrial organization models of third-degree price discrimination suggest that arbitrage generally improves welfare, as long as the incumbent does not significantly restrict supply in response to arbitrage. We show that these models cannot be applied to the electricity sector, as they do not allow for production-cost differences across regions in combination with limited transportation capacity. Marginal production costs for electricity vary greatly across regions because of political constraints (for instance, acceptance of nuclear power plants), geographical constraints (wind- and water power production) and differences in demand-side characteristics.[1] Moreover, transmission lines connecting electricity regions are often

[1]Demand affects marginal production costs in two ways. Demand characteristics determine the equilibrium portfolio of production plants in a region. Regions might have demand peaks at

congested, as they were not designed to handle commercial trading activities, but to transport emergency power.[2] Once these aspects are correctly taken into account, arbitrage is more likely to decrease welfare compared to the outcome in the standard price-discrimination model.

In the electricity market, transmission capacity allows for two functions: it enables both price arbitrage by consumers and production-cost minimization by the monopolist. It is the interaction of these two (sometimes conflicting) functions of transmission capacity which have led to the counter-intuitive welfare results of our paper.

In the absence of price arbitrage, the monopolist will use the entire transmission capacity and shift as much of its production as possible to the low-cost region. Production costs are minimized, and the monopolist uses regional price discrimination. As there is no competition for accessing the transmission line, the price for transmission is zero.

In the presence of arbitrage, arbitrageurs trade electricity from the low price- to the high-price region. In the high-price region, the incumbent generator loses market share to the arbitrageurs and therefore lowers the price. In the low-price region, he gains market size and raises the price. Hence, arbitrage reduces the price differential.

Apart from a possible effect on total supply, the welfare effect of a reduction of the regional price difference consists of two parts: it improves allocational efficiency among consumers, and it decreases production efficiency, as relatively more will be produced in the high-cost region. We show that it is likely that the negative effect outweighs the positive one. The introduction of market mechanisms for the

different moments in time, which means that regions might be operating on different parts of their supply function. Marginal production costs range from 0 EUR/MWh for nuclear power plants and some water plants, to 30 EUR/MWh for gas and coal plants, and more than 100 EUR/MWh for peak-power plants (i.e. plants with low fixed costs and high marginal costs).

[2]This can be illustrated for the UCTE system, the world largest synchronously interconnected electricity system, covering 23 European countries. Of the 39 cross-border connections, 24 are congested more than 75% of the time, and only five connections are never congested.

allocation of essential facilities (e.g. auctioning) therefore makes sense only in the following situations: (1) when sufficient investment is made in transmission capacities or (2) when the market power of the incumbent is broken in at least one of the two markets. If transmission capacity is sufficiently large, then regional cost differences do not matter, and arbitrage eliminates the incumbent's ability to price discriminate. Furthermore, more competitive markets eliminate the conflict between allocational and production efficiency, and arbitrage improves both.

The results of this paper are highly relevant for evaluating the market-coupling projects in Europe. These projects are meant to further integrate electricity markets and to improve cross-border arbitrage. For instance, the Dutch and the Belgian power exchanges have been coupled since November 2006, improving arbitrage between the countries. The stylized model we develop in this paper fits this market coupling nicely: *Production costs are very different* between the two countries due to differences in past energy policies: Belgium produces 55% nuclear, while the Netherlands relies more on gas production (it is a gas exporter) and combined heat- and power generation. The *transmission capacity between Belgium and the Netherlands is limited* and congested about 30% of the time. *One firm is in a dominant position* both in Belgium and in the Netherlands: Electrabel has a market share of about 80% in Belgium, and 20% in the Netherlands. The results of our paper predict that when the transmission line is congested (30% of the time), total welfare is reduced.[3]
The outline of the paper is as follows: Section 2 reviews the literature on third-degree price discrimination. In Section 3, we present a simple example highlighting the role of production-cost differences. Section 4 describes the model and explains how transmission capacity is allocated with and without arbitrage. Section 5 discusses, for different allocation mechanisms, the behavior of the monopolist. The

[3]This is the case under the assumption that the price-cost markup is larger in Belgium than in the Netherlands, a reasonable assumption given the fact that the incumbent generator in Belgium has a quasi monopoly.

paper concludes with a discussion of the welfare effects of arbitrage.

2 Literature review

This paper discusses regional price discrimination in electricity markets. Third-degree price discrimination occurs when a monopolist is able to charge different prices to different markets or groups of consumers for a homogeneous good. The standard price discrimination model, where trade is costless and production costs are uniform, has been studied extensively in the literature. In this section we review this literature and show that the results do not always hold for the electricity sector, which is characterized by interregional production-cost differences and limited transmission capacity.

One of the main insights in the literature (Tirole, 1988) is that the welfare effect of arbitrage is the combination of two effects: an allocational effect and an output effect. Costless arbitrage guarantees equal prices across regions as consumers take advantage of any interregional price difference. Therefore, trade among consumers results in an efficient allocation of the good, given that the marginal valuation of an additional unit is the same across consumers. This is the *allocational effect* of arbitrage. At the same time, the monopolist may react by increasing or decreasing total output under the effect of arbitrage. This is the *output effect*. A positive output effect increases welfare. The total welfare effect is the sum of both effects: arbitrage increases allocational efficiency, but at the same time it might induce a strategic response of the monopolist, thereby decreasing output efficiency.

For linear demand (and constant marginal production costs), Robinson (1933) shows that the output effect is zero. In this case, price discrimination should be forbidden, as output is not allocated efficiently. Also for the linear demand case, Layson (1988) shows graphically that price discrimination is most harmful for society when it is most profitable for the monopolist, as the welfare loss is proportional to the profit gain from price discrimination[4]. Results are less clear-cut for non-linear demand. However, if the output effect can be shown to be positive, then the welfare effect will be positive. Robinson (1933) shows that in the standard price-discrimination model the output effect of arbitrage depends on the curvature of the two demand curves. She shows that when demand is convex in the high-priced market and concave in the low-priced market, the output effect is positive. Also, if both curves are strictly concave, the output effect is positive when the low-priced market is more concave[5].

Arbitrage might increase welfare even when the output effect is negative. In the literature there are no simple guidelines as to the total welfare effect of arbitrage. Total welfare effects of price discrimination (with no distinction made between the output- and allocational effect) were studied by Varian (1985), who derives upper and lower bounds of the effect of price discrimination in terms of changes in market prices and output. Malueg (1993) quantifies the relative size of the welfare change caused by third-degree price discrimination. For concave demand functions, he shows, for instance, that with price discrimination, welfare will never decrease more than 33% and never increase more than 150%.

Our paper shows that the results of the standard literature on third-degree price discrimination cannot always be applied to the electricity sector. The differences and similarities with the classical model are the following:

[4]Schmalensee (1981) generalizes this result for n independent markets with arbitrary demand function curvatures and constant marginal costs. He shows that a necessary but not sufficient condition for social welfare-improving price-discrimination is that total output increases compared to the non-discriminatory situation. In particular, prohibiting price discrimination is always welfare-increasing for linear demand, as total quantity remains unchanged in both regimes.

[5]Formally, the low-priced region has, in absolute terms, a larger 'adjusted concavity' E,

$$E = -\frac{qp''(q)}{p'(q)}$$

where $p(q)$ is the inverse demand function. Robinson's criterion for determining the output effect is derived for infinitesimally small price changes. For large price changes, Shih et al. (1988) derive more general conditions for the sign of the total output effect.

We demonstrate that the welfare effect of arbitrage is the combination of *three factors*. As in the standard model, there is an allocational effect and an output effect, but there is also a new effect: the production efficiency effect, which requires that the goods are produced in the low-cost region. Arbitrage will typically improve allocational efficiency, but will reduce production efficiency.

With linear demand functions, we derive that the output effect of arbitrage is zero, but that *arbitrage will not always increase welfare*. Arbitrage decreases welfare when production cost differences are larger than the difference in the consumers' willingness to pay (the regional price difference).

With concave demand functions, adjusted concavity determines the output effect. For transmission capacities close to zero, arbitrage increases total output when the demand in the low-price region is more concave than in the high-price region, where concavity is measured as in Robinson (1933). Hence, the sign of the output effect is identical to that in the standard third-degree price-discrimination model.

Finally, instead of defining upper- and lower bounds for welfare changes, we derive *sufficient conditions for a positive welfare effect of arbitrage*. These conditions depend on a combination of the curvature of the demand functions (which determines the output effect), the elasticity of the demand functions (which determines the regional deadweight loss) and the price-cost margin (which links allocational and production efficiency).

The discussion in our paper is linked to two strands of literature. The first concerns the presence of imperfect arbitrage in models with price discrimination. The second deals with the abuse of market power in electricity markets.

Several authors have introduced imperfect arbitrage in price-discrimination models. When there are transaction costs or when goods are not perfectly homogeneous, arbitrage might lead to "leakage" of products from one market to another without eliminating the price difference completely. The monopolist can charge different prices in both regions, but if there is a price difference, then some of his production will leak from the low-priced region to the high-priced region. In this context, Varian (1985) derives a general model where sales by the monopolist might depend on the prices charged in both regions. Wright (1993) looks at a special type of imperfect arbitrage: arbitrageurs have to pay a fixed arbitrage cost. Ahmadi and Yang (2000) look at a model where arbitrage is imperfect because consumers value the sales of the monopolist (the authorized seller) higher than they value the goods from the arbitrageurs (unauthorized re-seller, parallel importer). This could be the result of different packaging or warranty conditions, for example. They show that in that case, it might be profitable for the monopolist to have some arbitrage, as it helps him to price discriminate consumers on the basis of their valuation of the (perceived) quality of the goods. Our paper is different from this literature, as we look at imperfect arbitrage that is caused by limited transmission capacity combined with regional production-cost differences.

The current paper contributes to the discussion in the electricity sector on the interaction between transmission and energy markets. Joskow and Tirole (2000) and Gilbert et al. (2003) model the microstructure of the transmission-rights market. Assuming that the transmission line is always congested, they show that the auction design determines whether arbitrage is perfect or not. Each type of auction therefore has a different impact on welfare. The focus of our current paper is different. While they assume generators to be located at one end of the line, we assume production capacity at both ends, which allows us to understand the effect of production-cost differences. We do not study auction design, however, as we consider only two extreme cases namely, perfect arbitrage and no arbitrage. Borenstein et al. (2000) discuss a Cournot generation duopoly. They assume that each player has production in one of the regions, and that arbitrage is perfect. Insufficient transmission capacity decreases the competition in electricity market. Our paper is different, as we study the impact of arbitrage.

A long-term version of our model, where new transmission capacity can be built at a fixed long-term marginal investment cost, is studied by Willems (2004).

3 Numerical example

A well-known result on third-degree price discrimination is that for linear demand functions, arbitrage always increases welfare, as long as both markets are served (see section 2). Before formally introducing our model, we show with a simple numerical example that this is no longer the case when there are production-cost differences between the regions and when transmission capacity is limited.

We consider two regions $i \in \{1, 2\}$. In each region there are price-taking consumers, represented by a linear demand function $q_i(p)$:

$$\begin{aligned} q_1(p) &= 8 - p \\ q_2(p) &= 6 - p \end{aligned}$$

In the standard third-degree price-discrimination model, the incumbent monopolist can sell freely in both markets and has production costs normalized to zero.

If there is no arbitrage, the monopolist will set the local monopoly prices in each region $p_1^{NA} = 4$ and $p_2^{NA} = 3$, i.e. the prices that maximize local profit $q(p_i) \cdot (p_i - 0)$. If there is arbitrage, then the monopolist sets a uniform price p for both regions. He maximizes the joint profit $(q_1(p) + q_2(p)) \cdot (p - 0)$, and sets the price $p = p_1^A = p_2^A = 3.5$. As arbitrage increases the allocative efficiency in the market, and total output remains constant $q_1 + q_2 = 7$, it is obvious that arbitrage increases total welfare, which can also be seen numerically. Welfare increases from 37.5 to 37.75.

Let us assume now that there are production-cost differences. The cost of production in region 1 is $\Delta c = 3$ and that in region 2 remains normalized to zero. Further, we assume that the transmission capacity between the two regions is limited to a total capacity of $k = 2$, but that there are no capacity limits on the production itself.

Without arbitrage, the monopolist will use the transmission line to import energy from the low-cost area to the high-cost area, up to the transmission capacity. By transporting goods from the low-cost area to the high-cost area, he reduces production costs with $k \cdot \Delta c$. The monopolist will use the transmission line up to capacity. The monopolist will set the local monopoly prices in each region, taking into account local production costs. He will maximize profit in the low-cost area $q_2(p_2) \cdot (p_2 - 0)$ by setting a price $p_2^{NA} = 3$ as before. In the high-cost region, the monopolist maximizes $q_1(p_1) \cdot (p_1 - \triangle c)$, by setting a price $p_1^{NA} = 5.5$. Total output is equal to $q_1 + q_2 = 5.5$.

With arbitrage, arbitrageurs will export energy from the low-priced region to the high-priced region. The monopolist will therefore sell less in the high-priced region, and more in the low-priced region. While setting prices, the monopolist takes this into account. In the high-priced area the monopolist sets the price $p_1^A = 4.5$, which maximizes his local profit $(q(p_1) - k)(p_1 - \Delta c)$. In the low-priced area, the monopolist sets the price $p_2^A = 4$, which maximizes local profit $(q(p_2) + k)(p_2 - 0)$.

Arbitrage decreases the price difference between the regions, but does not eliminate the difference completely. As before, there is no output effect, and total output is equal to 5.5 units with or without arbitrage.

As total output has remained constant, can we still conclude that arbitrage increases welfare? A quick calculation shows that this is not the case. Welfare $W = U_1(q_1) + U_2(q_2) - \Delta c \cdot (q_1 - k)$ decreases with arbitrage from 28.875 to 27.375, where $U_i(q_i)$ represents the gross consumer surplus. The reason is that arbitrage did not only change the allocation of the goods, but also the production location. Total production costs $\Delta c \cdot (q_1 - k)$ increased from 0.5 to 1.5.

In summary, the output effect of arbitrage is zero, and while arbitrage increased allocational efficiency (the regional price difference decreased), it reduced production efficiency. This latter effect outweighed the former effect, given that total welfare decreased under arbitrage.

4 Model description

This section presents a formal model on the effect of arbitrage in the electricity sector. It extends the standard third-degree price discrimination model, assuming that interregional transmission capacity is small, and that each region has different production costs.[6]

We compare two access regimes for the allocation of transmission capacity. In the first regime, the no-arbitrage regime (NA), we assume that the incumbent remains the only user of the transmission line. Arbitrageurs find it difficult or impossible to buy transmission capacity to profit from arbitrage on regional price differences. This might happen, for instance, when the incumbent owns the transmission capacity and access to the transmission line is inadequately regulated, or when the incumbent sells power with a resale restriction, forbidding consumers to resell their electricity and thereby drying up the liquidity on the energy markets. In the second regime, the arbitrage regime (A), the monopolist has to share the transmission line with arbitrageurs. Arbitrageurs buy transmission capacity and trade electricity from the low-price- to the high-price region until the price for transmission equals the price difference between the two regions. This occurs when access to the transmission

[6]Transmission capacity is "small" when, independent of the access regime, the transmission line is congested. In the appendix we define "small" transmission capacity as a function of demand functions and production costs. Note, however, that we neglect production constraints, and assume constant marginal production costs.

line is auctioned efficiently and each region has a well-functioning energy market. Some regional power exchanges go even further by collaborating and setting up trading systems that *by design* eliminate arbitrage opportunities, given the technical constraints of the system. This is called market coupling in Europe.[7] The model assumes that the monopolist is a "first mover" that has three decision variables in each regime: price setting in region 1 and in region 2, and determining the amount of transmission it will use. The monopolist perfectly foresees how consumers and arbitrageurs (if they are present) will react to his decisions. Consumers and arbitrageurs are modeled as "price takers" (i.e. they react to prices). Consumers decide, given the market price, how much energy they will consume. If arbitrageurs are present, they trade transmission capacity, until the price for transmission is equal to the price difference between the regions.

We assume that both the energy market and the transmission market clear simultaneously, and therefore do not model the micro-structure of the electrical energy- and the transmission markets. In particular, we will not describe the auction mechanism that is used to allocate transmission capacity. See Joskow and Tirole (2000) for a discussion of different mechanisms.

We choose to present the model assuming that the monopolist sets the regional price for electricity. Alternatively, we could build a model where the monopolist has three different decision variables: setting the amount of electricity it sells in each region and determining how much it will transport; however, the results would not change.[8] The advantage of our approach (prices are strategic variables) is that

[7]Note that Arbitrage (A) and No-arbitrage (NA) are two extreme cases. Arbitrage could be hampered even when transmission capacity is auctioned. This might happen if "gate closure", the moment when final bids have to be submitted in the auction, are different for the transmission market and the two power markets. If arbitrageurs need to buy transmission capacity before the electricity price is known, their risk increases, and arbitrage becomes less easy.

[8]This result holds in almost all models where there is one strategic player and all other players are price takers. In a standard monopoly model, for instance, it does not matter whether price or quantity is the strategic variable.

it is very similar to the standard third-degree price discrimination literature, and we can focus on the extra effects of transmission constraints and production-cost differences. However, the behavior of the arbitrageurs might be easier to explain in the quantity model, and also second-order conditions are more easily checked in a quantity framework (see Appendix).

4.1 Formal model

Consider two regions $i \in \{1, 2\}$. In each region there are price-taking consumers, represented by a downward-sloping and concave demand function $q_i(p)$. The incumbent player is active in both markets and has marginal production cost c_H in region 1 and c_L in region 2 ($c_H - c_L = \Delta c > 0$). Transportation from region 2 to region 1 is costless[9], but limited by the thermal transmission constraint k of the transmission line that connects both regions. If the demand for transportation is larger than the capacity of the line, it becomes a scarce good with a positive price. The price of the transmission rights will be denoted by τ. Access to the transmission line is sold at a price τ and the monopolist and arbitrageurs buy x_M and x_A transmission rights with $x_M + x_A \leq k$. The monopolist maximizes profit by setting the price p_i in region i, and by transporting x_M from region 2 to region 1. When it takes its decisions it will foresee the reactions of arbitrageurs and consumers. Their reaction will determine the price for transmission τ and the amount of transmission rights arbitrageurs buy, x_A. The resulting price is different for the two access regimes:

In the access regime without arbitrage (NA), the monopolist is the sole user of transmission capacity ($x_A^{NA} = 0$). As there are no arbitrageurs, there is no upward pressure on the transmission price, which is therefore zero ($\tau^{NA} = 0$). As a consequence, the monopolist uses the entire transmission capacity to import cheap units from the low-cost region ($x_M = k$).

[9] We neglect the losses on the network.

In the access regime with arbitrage (A), arbitrageurs will trade energy until the price for transmission capacity is equal to the price difference ($\tau^A = \Delta p \equiv p_1 - p_2$), and will buy all transmission capacity that is left on the market by the monopolist ($x_A^A = k - x_M$).[10]

The profit of the monopolist is equal to the revenue from selling s_i minus the production cost of producing r_i in region i, minus the transmission cost:

$$\pi = \underbrace{s_1 p_1 + s_2 p_2}_{\text{Revenue}} - \underbrace{r_1 c_H + r_2 c_L}_{\text{Production cost}} - \underbrace{\tau^t x_M}_{\text{Transmission cost}} \tag{1}$$

In region i, the monopolist sells s_i at a price p_i. Sales in region 1 are provided by producing r_1 locally and by importing x_M units from region 2 ($s_1 = r_1 + x_M$). Sales in region 2 are equal to the production in region 2 minus the export to region 1 ($s_2 = r_2 - x_M$). In addition, the monopolist needs to pay $\tau^t x_M$ for obtaining the transmission rights.

The sales s_1 in region 1 are equal to the demand in region 1 minus the amount that arbitrageurs import into region 1.

$$s_1 = q_1(p_1) - x_A \tag{2}$$

The sales s_2 in region 2 are equal to the demand in region 2 plus the amount that arbitrageurs export from region 2.

$$s_2 = q_2(p_2) + x_A \tag{3}$$

Equations 2 and 3 describe the sales of the incumbent. In the absence of arbitrage

[10] To be more precise, this result is valid only when $\Delta p \geq 0$. If the monopolist sets a negative price difference ($\Delta p \leq 0$), arbitrageurs would trade in the opposite direction and $x_A^A = -k - x_M \leq 0$. See appendix.

($x_A = 0$), the market "sealed". The monopolist could set any price in the two markets, without having leakage from one market to the other. With arbitrage, however, some "leakage" will occur, since arbitrageurs will buy electricity in the high-price region and sell it in the low-price region. This formulation is similar to Varian (1985).

Rewriting the monopolist's profit in equation 1, by assuming binding transmission capacity ($x_M + x_A = k$), and taking relations 2 and 3 into account, we obtain

$$\pi = q_1(p_1)(p_1 - c_H) + q_2(p_2)(p_2 - c_L) + \underbrace{(x_M + x_A)\Delta c}_{\text{production}} - \underbrace{x_A \Delta p}_{\text{leakage}} - \underbrace{x_M \tau}_{\text{transport}} \quad (4)$$

This formulation is similar to the classical third-degree discrimination model. However, it includes three extra terms. Producing goods in the low-cost region reduces total costs for the monopolist. Given the "leakage", the monopolist loses profitable sales in the high-priced region, and these losses cannot be made up by selling more in the low-priced region. The last term is the monopolist's cost of buying transmission rights.

5 The incumbent at work

This section derives the strategy of the incumbent under the two access regimes and describes how the monopolist adjusts his strategy in response to arbitrage. We show the following: The monopolist understands that with arbitrage, price discrimination is less profitable (equation 4), as leakage will occur ($x_A \geq 0$) and transportation will become costly ($\tau = \Delta p$). Therefore, the incumbent decides to reduce the interregional price difference, reducing both the leakage and transportation costs. We now discuss both pricing strategies in turn.

5.1 Exclusive use by the incumbent

In the first case, the incumbent has exclusive access to the line. The monopolist maximizes his profit (4) by choosing the prices p_i in region i and transporting the amount x_M. As transmission capacity is small, the monopolist will use all available capacity of the line to substitute expensive generation in region 1 with cheap generation in region 2 ($x_M = k$). The profit equation simplifies to

$$\pi^{NA} = q_1(p_1)(p_1 - c_H) + q_2(p_2)(p_2 - c_L) + k\Delta c. \tag{5}$$

The prices set by the monopolist p_1^{NA} and p_2^{NA} are determined by the standard inverse elasticity rule:

$$\frac{p_1^{NA} - c_H}{p_1^{NA}} = \frac{1}{\varepsilon_1} \tag{6}$$

$$\frac{p_2^{NA} - c_L}{p_2^{NA}} = \frac{1}{\varepsilon_2}; \tag{7}$$

with $\varepsilon_i = -p_i \frac{q_i'(p_i)}{q_i(p_i)}$ denoting the demand elasticity in region i. Hence, without arbitrage, regional prices are equal to the local monopoly prices, taking local production costs into account:

$$p_1^{NA} = p_1^M(c_H) \tag{8}$$

$$p_2^{NA} = p_2^M(c_L). \tag{9}$$

In the appendix, we describe the optimization problem 5 of the monopolist in more detail, and we show that equations 6 and 7 are indeed necessary and sufficient conditions for the prices set by the monopolist, under the assumption that transmission capacity is sufficiently small. This is not obvious, as the optimization problem of the monopolist is not convex in prices, and therefore first-order conditions are not sufficient to find a global optimum.

5.2 Market allocation of transmission

In the second case, arbitrageurs can obtain access to the line and buy transmission capacity. This case is denoted A (arbitrage). Again, the monopolist maximizes profit (4) by setting the price p_i and choosing x_M. In the case where there is arbitrage, the price for transmission τ is equal to the regional price difference Δp.

In the appendix, we show that the monopolist finds it in its own interest to set a positive price difference $\Delta p > 0$, as long as transmission capacity is sufficiently small. In order for this to be the case, we will show it is sufficient to assume that the monopoly price in the high-cost region is higher than the monopoly price in the low-cost region $p_1^M(c_H) > p_2^M(c_L)$. This assumption ensures that arbitrageurs and the incumbent have an incentive to trade in the same direction. The monopolist wants to transport energy from the low-cost region to the high-cost region, and the arbitrageurs from the low-price to the high-price region. The assumption that the monopoly price in the high-cost region is high, is valid when the demand function is concave and similar in both regions. If the price in the high-cost region is below the price in the low-cost region, then arbitrageurs will have the incentive to export energy from the high-cost region and import it to the low-cost region. It is obvious that this will increase production costs and that arbitrage is likely to decrease welfare. The interested reader can check Willems (2002) for a more detailed discussion of this situation.

Arbitrageurs will buy all remaining transmission capacity $x_A = k - x_M$ to arbitrage away price differences. Hence, the transmission capacity is binding ($x_M + x_A = k$). The profit of the monopolist can be rewritten as follows:

$$q_1(p_1)(p_1 - c_H) + q_2(p_2)(p_2 - c_L) + k(\Delta c - \tau). \tag{10}$$

Clearly, this objective function depends on p_1 and p_2. Changing the price p_i impacts not only the regional profits $q_i(p_i)(p_i - c_j)$ but also the total transmission cost $k\tau$.[11] The monopolist will set the prices p_1^A and p_2^A according to an *adjusted* standard inverse elasticity rule

$$\frac{p_2^A - c_L}{p_2^A} = \frac{1}{\varepsilon_2}\frac{r_2^A}{q_2^A} \tag{11}$$

$$\frac{p_1^A - c_H}{p_1^A} = \frac{1}{\varepsilon_1}\frac{r_1^A}{q_1^A}. \tag{12}$$

The relative price-cost margin is equal to the inverse of the elasticity multiplied by a correction factor: the ratio of local production r_i and local consumption q_i. In the appendix, we show that these first-order conditions are necessary and sufficient conditions for the global optimum of the monopolist, as long as transmission capacity is sufficiently small.

5.3 Comparison

This section compares the prices set by the monopolist under both access regimes, and shows that arbitrage will induce the monopolist to reduce the regional price difference, for a small transmission capacity $k \to 0$. The argument goes as follows: in the extreme case of zero transmission capacity ($k = 0$), local production is equal to local consumption ($r_i = q_i$ in equations 11 and 12). According to equations 6, 7, 11 and 12, the profit-maximizing prices are identical in both access regimes, and are

[11] Note that it does not matter who uses the transmission line. If the monopolist buys transmission capacity, he has to pay $x_M\tau$. If arbitrageurs buy transmission capacity, the monopolist will sell x_A units more in the low-price region, but lose x_A units in the high-price region. In total, this loss due to arbitrage is equal to $x_A\Delta p = x_A\tau$. Hence, the total cost of transmission amounts to $(x_M + x_A)\tau = k\tau$, regardless of whoever uses the line.

equal to the local monopoly prices:

$$p_1^A = p_1^{NA} = p_1^M(c_H) \tag{13}$$

$$p_2^A = p_2^{NA} = p_2^M(c_L). \tag{14}$$

A marginal increase of transmission capacity away from zero does not affect prices in the no-arbitrage regime. This is not the case when arbitrageurs are active, as

$$\frac{\partial p_2^A}{\partial k} = \frac{1}{q_2'}\frac{1}{E_2 - 2} \geq 0 \tag{15}$$

$$\frac{\partial p_1^A}{\partial k} = -\frac{1}{q_1'}\frac{1}{E_1 - 2} \leq 0, \tag{16}$$

with $E_i = -\frac{q_i p_i(q_i)''}{p_i(q_i)'}$ the relative curvature coefficient (or adjusted concavity) of the demand function in region i. Hence, for concave functions, the price in the low-price region 2 increases and the price in the high-price region 1 decreases when transmission capacity increases. If the transmission capacity is small, then price arbitrage reduces the price difference between the regions, compared to the case without arbitrage.

6 This paper's contributions

Highlighted in this section are three contributions made by this paper to the third-degree price discrimination literature. First of all, we prove that for the electricity market, the effect of arbitrage is no longer the combination of only an allocational and an output effect, but that also a third effect needs to be taken into account: the production effect. Then, contrary to the standard results, we show that for linear demand functions, arbitrage is often welfare decreasing, because production efficiency is reduced. It is the case when demand functions are similar and cost differences relatively large. Finally, we define necessary and sufficient conditions for welfare to increase under arbitrage for the category of concave demand functions.

6.1 The production effect

The main intuition from the literature (See section 2) on third-degree price discrimination is that the welfare effect of arbitrage can be determined as the sum of an allocational effect (which is positive) and an output effect (which can be positive or negative). Arbitrage improves allocational efficiency, as it allocates a given output more efficiently among consumers.[12] Arbitrage might have a positive or a negative output effect, as the monopolist might decrease or increase its output level in response to arbitrage (Tirole, 1988).

If we introduce transmission-capacity constraints and production-cost differences, this result is no longer valid, as a third effect - the production effect - needs to be considered. Arbitrage not only affects the allocation of goods among consumers, but also the location of production. Welfare decreases when electricity is produced less efficiently by shifting production from the low-cost to the high-cost region. Hence, arbitrage has three effects: an output effect, an allocational effect and a production effect. However, as long as the transmission line remains congested, the allocational and the production effects are linked with each other. The production and consumption locations cannot be chosen independently. Therefore, we define the interregional effect as the combination of these two effects. Typically, there is a trade-off: arbitrage increases allocational efficiency but reduces production efficiency. Interregional efficiency is achieved when the two effects are balanced: the gain of consumer surplus is equal to the loss of production efficiency. The welfare effect of arbitrage is the sum of an output effect and the interregional effect, where the latter is the combination

[12] Without arbitrage, consumers in one region have a higher willingness to pay than do their counterparts in the other region, and a reallocation of goods would lead to a Pareto improvement.

of the allocational and production effects.

Define social welfare W as the sum of consumer surplus minus production costs, which can be written as a function of quantities q_1, q_2 and transport $x = x_A + x_M$:

$$W(q_1, q_2, x) = U_1(q_1) - q_1 c_H + U_2(q_2) - q_2 c_L + x\Delta c$$

We say that arbitrage increases interregional efficiency if (keeping total production $q = q_1 + q_2$ constant) welfare increases with a decrease of the price difference Δp. Mathematically, this is defined as the partial derivative of welfare with respect to the price difference:

$$\left.\frac{\partial W}{\partial(\Delta p)}\right|_{dq=dx=0} \leq 0 \tag{17}$$

Note that this partial derivative is taken under the assumption that the transmission line remains congested $x = k$.

Proposition 1 *If transmission capacity is fully used and if the price difference is smaller than the production-cost difference $\Delta p \leq \Delta c$, then arbitrage decreases interregional efficiency.*

Proof. We are interested in the marginal welfare effect dW

$$dW = (p_1 - c_H)dq_1 + (p_2 - c_L)dq_2 + \Delta c \cdot dx \tag{18}$$

in response to a change of the price difference $d(\Delta p)$

$$d(\Delta p) = \frac{dp_1}{dq_1}dq_1 - \frac{dp_2}{dq_2}dq_2 \tag{19}$$

when the transmission line is congested $dx = 0$, and total production remains constant

$$dq = dq_1 + dq_2 = 0. \tag{20}$$

Combining equations 18, 19 and 20, we find that

$$\left.\frac{\partial W}{\partial(\Delta p)}\right|_{dq=dx=0} = \rho \cdot [\Delta c - \Delta p] \, , \tag{21}$$

with $\rho^{-1} = -(q_1'^{-1} + q_2'^{-1}) > 0$.

The equation shows that the welfare effect of a marginal change of the price difference is proportional to $(\Delta c - \Delta p)$. If ρ units of demand are shifted from region 1 to region 2, then also ρ units of production need to be shifted from region 1 to region 2, as the transmission capacity is assumed to be binding. At the margin, shifting demand decreases consumer surplus by $\rho\Delta p$, and decreases production costs by $\rho\Delta c$. Combining equation 21 with the result of section 5.3 (arbitrage reduces the regional price difference Δp) proves the proposition. ■

Proposition 1 shows that if there are regional cost differences and limited transmission capacity, then interregional efficiency depends not only on the location of consumption but also on the location of production. Although arbitrage reduces the price difference, which improves the allocation of consumption, it may also worsen the allocation of production. Depending on regional price- and cost differences, the effect of arbitrage on interregional efficiency might be positive or negative. In the optimum, demand and production should be allocated such that the price difference equals the cost difference.

We can apply the intuition of this result to the Belgian-Dutch electricity market. Electrabel has a dominant position in Belgium (90%) and in the Netherlands (20%). With respect to the model, Belgium is the low-priced country that exports energy to the Netherlands ($p_{BE} < p_{NL}$). The transmission line is congested about 30% of the time.[13] In November 2006 the two electricity markets were coupled, in order to improve cross-border arbitrage between the two countries. The two countries might

[13]Data provided by the Dutch transmission system operator, TenneT.

have substantial production-cost differences, as electricity generation is essentially nuclear (55%) and gas fired (28%) in Belgium, whereas it is mainly gas (64%) and coal fired (23%) in the Netherlands.[14] Assuming that Electrabel's competitors in the Netherlands behave as a competitive fringe[15], the welfare effects of improved arbitrage can be predicted by our model. Welfare will decrease when the price difference is smaller than the production-cost difference (and the output effect is negligible). It is not straightforward to determine the marginal cost in the two countries, as this depends not only on the generation mix but also on other factors such as the time of day, environmental regulation, the maintenance schedule, and the price of fuel and CO_2 permits. We therefore do not derive the welfare effect directly (by comparing price and cost differences for every possible configuration) but indirectly (by looking at the incentives of Electrabel). Electrabel has fewer incentives to increase prices above the marginal cost in the Netherlands than it does in Belgium, as Electrabel's residual demand in the Netherlands is flatter than in Belgium (given competition with the fringe), and as its sales in the Netherlands are smaller than in Belgium. Price-cost mark-ups are therefore higher in Belgium than in the Netherlands:

$p_{BE} - c_{BE} > p_{NL} - c_{NL}$.[16] Hence, when the transmission line is congested, then the cost difference between the Netherlands and Belgium is larger than the price difference, $c_{NL} - c_{BE} > p_{NL} - p_{BE} > 0$, and arbitrage will reduce welfare (if the output effect is small).[17] In periods in which there is no congestion (70% of the time),

[14] Eurostat.

[15] Nuon, Essent and RWE each have 20% of the capacity in the Netherlands, and smaller generators own the remaining 20%.

[16] The price-cost markup in region i depends on the slope of the residual demand function and the total sales in a region: $p_i - c_i = -q_i p^R{}'(q_i)$ where $p^R{}'(q_i)$ is the slope of the inverse residual demand function. Electrabel sells less in the Netherlands ($q_{NL} < q_{BE}$), and faces a flatter residual demand function$|p^R_{NL}{}'(q_{NL})| < |p^R_{BE}{}'(q_{BE})|$, hence mark-ups are smaller.

[17] It is not straightforward to determine the marginal cost in the two countries, as this depends not only on the generation mix but also on other factors such as the time of day, environmental regulation, the maintenance schedule, and the price of fuel and CO_2 permits.

arbitrage will increase welfare by reducing the possibility to price discriminate. The overall effect of arbitrage could go either way.

6.2 Welfare effects for linear demand

This section examines the welfare effects of arbitrage when the demand functions are linear and take the form $q_i(p) = \alpha_i - \beta_i p$ with $\alpha_i > 0$ and $\beta_i > 0$. We use the linear model to illustrate the proposition we derived in the previous section. The advantage of using linear functions is that the output effect is zero (Robinson, 1933), which allows us to concentrate on the interregional effects of arbitrage. Recall from section 6.1 that interregional efficiency requires the price difference to be equal to the cost difference between the two regions $\Delta p^{opt} = \Delta c$.

Straightforward calculation shows that the price difference under arbitrage (A) and no arbitrage (NA) is equal to

$$\Delta p^{NA} = \frac{\Delta c}{2} + \frac{\chi}{2} \tag{22}$$

$$\Delta p^{A} = \frac{\Delta c}{2} + \frac{\chi}{2} - \frac{k}{2}\left(\frac{1}{\beta_1} + \frac{1}{\beta_2}\right), \tag{23}$$

where $\chi = \frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}$ is a measure of the regional difference of consumer preferences. When the two demand functions are similar ($0 \leq \chi < \Delta c$), then the price difference without arbitrage is below the optimum ($\Delta p^{NA} \leq \Delta c$). The intuition for this is that the monopolist will not fully pass along an increase in production costs to consumers. As arbitrage makes price discrimination costly for the monopolist, the monopolist will react by decreasing the price difference ($\Delta p^{NA} \geq \Delta p^{A}$). Therefore, as the price difference was already too low, arbitrage lowers the interregional efficiency

$$\Delta c = \Delta p^{opt} > \Delta p^{NA} > \Delta p^{A} \tag{24}$$

and hence (as there is not output effect) lowers welfare.[18] This is summarized in the following proposition:

Proposition 2 *For similar linear demand functions ($0 \leq \chi < \Delta c$), and with binding transmission constraints, arbitrage is welfare decreasing.*

Proof. The proof follows from the discussion above. ■

In contrast to the standard literature (Robinson, 1933), we have shown that arbitrage decreases welfare for linear demand functions when consumers have similar preferences, production-cost differences matter and transmission capacity is binding.

The following corollary shows that allowing for arbitrage makes sense only when the market power of the incumbent is broken in at least one of the two regions:

Corollary 3 *If there is perfect competition in the low-cost region 2 and linear demand in both regions, then arbitrage is always beneficial.*

Proof. If there is perfect competition in the low-cost region 2, then the residual demand function for the monopolist is perfectly elastic $\beta_2 \to \infty$, and the price in the low-cost region is $\frac{\alpha_2}{\beta_2} = c_L$. The price in the high-cost region 1 is always above c_H. Arbitrage gives an incentive to decrease the price difference between the regions, and hence decreases the price in region 1, which is always optimal. ■

Practically, this means that market coupling between, for instance, the Netherlands (with a relatively concentrated and high-cost market) and Norway (with a competitive and low-cost market) is not likely to raise competition policy issues.

[18] It might be instructive to look back at the results of section 3. In the example, demand functions are relatively similar, as $\chi = 8 - 6 = 2 \leq \Delta c = 3$. Arbitrage does not change total production, but reduces the price difference from $\Delta p^{NA} = 2.5$ to $\Delta p^{A} = 0.5$. However, the optimal price difference is $\Delta c = 3$; hence, total welfare decreased.

6.3 Welfare and output effect for non-linear demand

When demand is non-linear, the monopolist will not only change the price difference in response to arbitrage, but will also adjust total production. This section derives the output- and welfare effects of arbitrage for concave demand functions and for small transmission capacities.

We show that the output effect depends crucially on the curvature of the demand functions and obtain similar results as in the standard model on third-degree price discrimination (Robinson, 1933).

We further explain that in order to study the welfare effect of arbitrage, one needs to compare the regional demand functions with respect to three factors: the curvature of demand (which determines the output effect), the regional price level (which determines the allocational effect), and the elasticity of demand (which determines the price-cost margin, and therefore (indirectly) the production effect).

In order to derive our results, we assume that transmission capacity is "infinitesimally" small and evaluate total output (and total welfare) under the two regimes using a Taylor expansion around $k = 0$. We use the fact that the two regimes give identical results when $k = 0$. Our results are in that sense similar to Robinson (1933), who assumes that price differences are (infinitesimally) small. We do not derive results for larger transmission capacities, as this would require integration of the output (and welfare) functions over a range of different levels of k, and, in order to make general statements when comparing these integrated functions, one would need to make assumptions on even higher order derivatives of the demand functions. This would not provide any additional intuition.

The section starts with the output effect, followed by the welfare effect.

Define the total demand level $Q^l(k)$ under regime $l = A, NA$ with transmission capacity k as

$$Q^l(k) \equiv q_1^l(k) + q_2^l(k),$$

where $q_i^l(k) \equiv q_i(p_i^l(k))$ is the demand in region i when the monopolist sets prices $p_i^l(k)$ under regime l.

The following proposition gives necessary and sufficient conditions for the output effect to be positive $Q^A(k) \geq Q^{NA}(k)$, and derives a similar result as Robinson (1933) .

Proposition 4 *Arbitrage increases total output for transmission capacities close to zero* ($k \to 0$) *if and only if demand in the low-cost region is more concave, i.e. when*

$$E_1(p_1^m(c_H)) \geq E_2(p_2^m(c_L)), \tag{25}$$

where $E_i = -\frac{q_i p_i''(q_i)}{p_i'(q_i)} = \frac{q_i q_i''(p_i)}{(q_i'(p_i))^2} < 0$ *denotes the adjusted concavity.*

Proof. Using the Taylor approximation for small $k \to 0$, arbitrage increases total production if and only if

$$Q^A(0) + k \cdot \frac{dQ^A}{dk}(0) \geq Q^{NA}(0) + k \cdot \frac{dQ^{NA}}{dk}(0).$$

This expression can be simplified. When transmission capacity $k = 0$, then the total level of consumption does not depend on the access regime $Q^A(0) = Q^{NA}(0)$. Without arbitrage, increasing the size of the transmission capacity does not change total production (hence, $\frac{dQ^{NA}}{dk}(0) = 0$). Therefore, arbitrage increases total output if and only if increasing the size of the transmission line would increase output under the arbitrage regime

$$\frac{dQ^A}{dk}(0) \geq 0.$$

Using equations 15 and 16, we can derive the effect of transmission capacity on total demand when there is arbitrage:

$$\begin{aligned}\frac{dQ^A}{dk}(0) &= \frac{dq_1^A}{dk}+\frac{dq_2^A}{dk}\\ &= \frac{1}{2-E_1(p_1^A)}-\frac{1}{2-E_2(p_2^A)}.\end{aligned}$$

Rearranging this expression gives the proof of the proposition. ■

In sections 6.1 and 6.2 we showed that arbitrage might *decrease* interregional efficiency in the market and that an increase of total output is therefore *no longer a sufficient condition for welfare to increase* under arbitrage. Arbitrage will increase welfare only when a large increase in output offsets the reduction of interregional efficiency.

Define the resulting level of welfare V under regime $l = A, NA$ as

$$V^l(k) = W(q_1^l(k), q_2^l(k), k) \quad l = A, NA. \tag{26}$$

The following proposition derives necessary and sufficient conditions for the welfare effect to be positive $V^A(k) \geq V^{NA}(k)$ for small transmission capacities. As in Varian (1985), we derive conditions for the welfare effect, without relying on the output effect.

Proposition 5 *Define the function* $f_i(p) = \frac{\varepsilon_i(p)}{p}(2 - E_i(p))$*, with* $\varepsilon_i(p) = -\frac{q_i'(p_i)}{q_i(p_i)}p_i$ *as the demand elasticity in region* i*. Arbitrage increases welfare for transmission capacities close to zero* ($k \to 0$) *if and only if*

$$f_2(p_2^m(c_L)) > f_1(p_1^m(c_H)). \tag{27}$$

Proof. Arbitrage increases welfare for small transmission capacities when

$$V^A(0) + k \cdot \frac{dV^A}{dk}(0) \geq V^{NA}(0) + k \cdot \frac{dV^{NA}}{dk}(0). \tag{28}$$

Here, we make a first-order approximation of the welfare function V around $k = 0$, taking into account the behavior of the monopolist. Welfare is equal under both regimes for $k = 0$ ($V^A(0) = V^{NA}(0)$), so that both terms drop out of the inequality. The marginal welfare effect of transmission capacity can be calculated as

$$\frac{dV^l}{dk} = (p_1^l - c_H)\frac{dq_1^l}{dk} + (p_2^l - c_L)\frac{dq_2^l}{dk} + \Delta c \frac{dx^l}{dk}.$$

It is the sum of three parts: The final term in the expression is the cost advantage of extra transmission capacity: production in the high-cost region is substituted by low-cost production. The first two terms describe the effect of transmission capacity on welfare in each of the regions, taking into account the adjustments of the incumbent in response to an increase of transmission capacity.

Without arbitrage, the marginal effect of transmission capacity is

$$\frac{dV^{NA}(0)}{dk} = \Delta c > 0. \tag{29}$$

With larger transmission capacity, electricity will be produced more cheaply, thereby increasing production efficiency, although prices and, hence, consumption remain constant (the first two terms are zero).

With arbitrage, consumption in the two regions depends on the size of the transmission line, as the monopolist will reduce the price difference between the two regions. The marginal effect of transmission capacity on welfare is equal to

$$\frac{dV^A(0)}{dk} = \frac{p_1^A - c_H}{2 - E_1(p_1^A)} - \frac{p_2^A - c_L}{2 - E_2(p_2^A)} + \Delta c. \tag{30}$$

Increasing transmission capacity changes regional welfare levels (the first two terms are different from zero) and allows for more export from the low-cost region. Note that the cost advantage from importing more from the low-cost region is identical in the cases with and without arbitrage. Hence, using equations 29 and 30, arbitrage is welfare-improving if and only if

$$\frac{\varepsilon_1^A}{p_1^A}(2 - E_1(p_1^A)) \leq \frac{\varepsilon_2^A}{p_2^A}(2 - E_2(p_2^A)). \tag{31}$$

Here we also use the relation between the relative price-cost margin and the demand elasticity as given by the first-order conditions 11 and 12. This proves the proposition. ■

If the low-cost region is more elastic, more concave and has a lower monopoly price, then arbitrage increases welfare. This generalizes corollary 3: if the low-cost region is perfectly competitive ($\varepsilon = +\infty$ and$E = 0$), then arbitrage will always increase welfare.

For equal regional demand functions, the following corollary immediately follows:

Corollary 6 *If two regions have identical concave demand functions* $q_i(\cdot) = \widetilde{q}(\cdot)$ *and* $f' < 0$ (> 0) *for all* p*, then arbitrage increases (decreases) welfare.*

Proof. If both regions have an identical concave demand function: $q_1(\cdot) = q_2(\cdot) = q(\cdot)$, then the monopolist passes cost differences only partially along to his consumers $0 < \frac{\partial p^M}{\partial c} < 1$. The monopolist sets the highest price in the high-cost region: $p^M(c_L) < p^M(c_H)$. Using the fact that $f'(p) > 0$, the results follow directly from proposition 5. ■

In order to illustrate the results of this corollary, assume identical quadratic demand functions $q(p) = \alpha - \beta p - \gamma p^2$ for both regions $(\alpha, \beta, \gamma \in \Re^+)$. Arbitrage will reduce welfare as long as the coefficient γ of the second-order term is sufficiently small ($\gamma < \frac{1+\sqrt{3}}{2}\frac{\beta^2}{\alpha}$), which ensures that $f'(p) > 0$ for all $p \geq 0$.[19] This result for quadratic functions naturally extends the linear demand case where $\gamma = 0$. For large values of γ, arbitrage might increase or decrease welfare, and no general conditions can be formulated.

7 Conclusion

This paper analyzes third-degree price discrimination in electricity markets. It develops a model in which a monopolist sells a final good in two regional markets, transmission capacity between regions is limited and production costs are high in one of the regions. We investigate whether access to the transmission capacity should be auctioned in order to facilitate regional arbitrage or whether the dominant incumbent firm should have the exclusive usage rights.

We show that the results of the classical third-degree price-discrimination model do not hold for electricity markets. The classical model identifies two effects of arbitrage on social welfare: the allocational effect and the output effect. Allocational efficiency requires that marginal willingness to pay be equalized across regions. Therefore, for a given total output, arbitrage ensures an efficient allocation of this output as regional price differences disappear. On the other hand, total output may increase or decrease as a result of arbitrage, where an increase is associated with higher social welfare. For electricity markets we identify a third welfare effect of arbitrage: production efficiency. *Arbitrage will typically reduce production efficiency*, as less electricity is produced in the low-cost region.

With regard to linear demand, the classical model shows that arbitrage increases

[19]For the quadratic demand function, $f'(p) > 0$ if and only if $E(p) > -1 - \sqrt{3}$. Arbitrage will certainly reduce welfare if this condition holds for all $p \geq 0$.

welfare, as total output remains constant, but that the reduction of regional price differences increases allocational efficiency. In electricity markets, arbitrage will reduce welfare if cost differences are important.

Regarding concave demand functions, the paper then derives necessary and sufficient conditions for welfare to increase with arbitrage. It is shown that arbitrage increases welfare when the low-cost region has a more elastic and less concave demand function and a lower autarky price.

The policy recommendation of this paper is that the introduction of market mechanisms for the allocation of essential facilities only makes sense when sufficient investment is made in transmission capacities or when the market power of the incumbent is broken in at least one of the two markets.

References

[1] **Ahmadi, Reza, and Rachel B. Yang.** 2000. "Parallel imports: Challenges from unauthorized distribution channels." *Marketing Science*, 19(3): 279-294.

[2] **Borenstein, Severin, James Bushnell, and Steven Stoft.** 2000. "The competitive effects of transmission capacity in a deregulated electricity industry." *RAND Journal of Economics* 31(2): 294-325.

[3] **Gilbert, Richard, Karsten Neuhoff, and David Newberry.** 2003. "Allocating transmission to mitigate market power in electricity networks." Cambridge University, CMI Working Paper 07.

[4] **Joskow, Paul L., and Jean Tirole.** 2000. "Transmission rights and market power on electric power networks." *RAND Journal of Economics*, 31(3): 450-487.

[5] **Layson, Stephen.** 1988. “Third-degree price discrimination, welfare and profits: A geometrical analysis.” *American Economic Review*, 78(5): 1131-1132.

[6] **Luo, Zhi-Quan, Jong-Shi Pang, and Daniel Ralph.** 1996. *Mathematical Programs with Equilibrium Constraints*. New York: Cambridge University Press.

[7] **Malueg, David A.** 1993. “Bounding the Welfare Effects of Third-Degree Price Discrimination.” *American Economic Review*, 83(4): 1011-1021.

[8] **Robinson, Joan.** 1933. *Economics of Imperfect Competition*. London: Macmillan.

[9] **Schmalensee, Richard.** 1981. “Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination.” *American Economic Review*, 71(1): 242-247.

[10] **Shih, Jun-ji, Jung-Chao Liu, and Chao-Cheng Mai.** 1988. “A General Analysis of the Output Effect Under Third-Degree Price Discrimination.” *Economic Journal*, 98(389): 149-158.

[11] **Tirole, Jean.** 1988. *The Theory of Industrial Organization*. Cambridge: The MIT press.

[12] **Varian, Hal R.** 1985. “Price Discrimination and Social Welfare.” *American Economic Review*, 75(4): 870-875.

[13] **Willems, Bert.** 2002. “Barring consumers from the electricity network might improve social welfare.” Katholieke Universiteit Leuven, CES-ETE, Working Paper Series, 2002-13.

[14] **Willems, Bert.** 2004. “Electricity Networks and Generation Market Power.” PhD diss. Katholieke Universiteit Leuven.

[15] **Wright, Donald J.** 1993. "Price-discrimination with transportation costs and arbitrage." *Economic Letters*, 41(4): 441-445.

A Appendix

This appendix describes the optimization problem of the monopolist and specifies the assumptions of the paper. Its main objective is to show that for sufficiently small transmission capacities, the equilibrium prices are given by equations 6, 7, 11 and 12. The model in the appendix is richer than that described in the main text, as we recognize that the monopolist might sometimes have the incentive to transport energy from the high-cost to the low-cost region (for reasons of price discrimination) and we take into account minimal production and consumption constraints. We show that the monopolist solves a non-convex optimization problem, and that first-order conditions are not always sufficient for a global optimum. The appendix contains four subsections. Subsection A.1 gives the assumptions and definitions of the model. We then solve the optimization problem of the monopolist without and with arbitrage. The last subsection combines the results of the first three subsections.

A.1 Assumptions and definitions

We describe demand in region $i = 1, 2$ by a concave demand function $q_i(p)$. Local unit production costs in regions 1 and 2 are respectively c_H and c_L with $\Delta c = c_H - c_L \geq 0$. For large levels of demand, the price in both regions is lower than the marginal cost: $\lim_{Q\to\infty} P_i(Q) \leq c_L$ where $P_i(Q)$ is the inverse demand function. This condition guarantees that the optimum is found at finite production capacities. In order to simplify notation, we denote by p_{ij}^M the price that a fictive monopolist with production cost c_j $j =\in \{L, H\}$ would set in region i :

$$p_{ij}^M \equiv \arg\max_p (q_i(p)(q_i - c_j)). \tag{32}$$

Using this definition, the autarky price in regions 1 and 2 (i.e. the monopoly price when there is no interregional trade) is p_{1H}^M and p_{2L}^M. One of the main assumptions of the paper is that the high-cost region has a larger autarky price than the low-cost region:

$$p_{1H}^M \geq p_{2L}^M$$

This assumption will reduce considerably the number of local equilibria we need to consider in the model. Further, we assume that in the standard arbitrage model without transmission constraints (but with identical demand specification), the monopolist will not find it profitable to "shut down" one of the markets and to sell all goods in the other market at that market's local monopoly price. A sufficient condition for this is that the reservation price in region 2 $\bar{p}_2 = P_2(0)$ is larger than the monopoly price in region 1 ($\bar{p}_2 \geq p_{1H}^M$). In this case, the monopolist will set a uniform price p_{totL} in both markets such that

$$\frac{p_{totL} - c_L}{p_{totL}} = \frac{1}{\varepsilon_{tot}}, \tag{33}$$

where ε_{tot} is the elasticity of total demand of the two regions taken together.

A.2 Optimization program without arbitrage

Without arbitrage (NA), the incumbent has the exclusive use of the transmission capacity and solves the following optimization problem:

$$\begin{aligned}
\max_{p_1, p_2, x} \pi^{NA} \quad &= \quad (p_1 - c_H)q_1(p_1) + (p_2 - c_L)q_2(p_2) + x\Delta c \\
&s.t. \quad q_1(p_1) \geq \max(0, x) \\
&\qquad q_2(p_2) \geq \max(0, -x) \\
&\qquad k \geq x \geq -k
\end{aligned}$$

The variable x represents the amount of electricity transported from region 2 to region 1. Negative numbers mean transport levels in the opposite direction. As production cannot become negative, consumption in a region should be larger than imports into that region. Furthermore, the amount of electricity transported from one region to the other cannot exceed the transmission capacity of the line k. Note that we allow prices to be negative.

The objective of the monopolist is not concave for low price levels, because the price in the high-cost region 1 might be below the marginal cost c_H when transmission capacity is sufficiently large. It is therefore by no means guaranteed that a solution of the first order conditions 6 and 7 gives a unique set of prices and transportation levels that maximize the profit of the monopolist. There might be several local maxima (and local minima) satisfying these conditions.

Second-order conditions In order to prove that the first-order conditions are necessary and sufficient for the global optimum, we rewrite the problem using quantities. The decision variables (p_1, p_2, x) of the optimization problem can be uniquely transformed to consumption and production quantities (Q_1, Q_2, r_1, r_2) by adding an energy conservation constraint.

$$\begin{aligned} \max_{Q_1,Q_2,r_1,r_2} \pi^{NA} \quad &= \quad P_1(Q_1)Q_1 + P_2(Q_2)Q_2 - c_H r_1 - c_L r_2 \\ & s.t. \qquad Q_i \geq 0 \; r_i \geq 0 \\ & \qquad\qquad Q_1 + Q_2 = r_1 + r_2 \\ & \qquad\qquad k \geq Q_1 - r_1 \geq -k \end{aligned}$$

This optimization problem has an (unbounded) convex set of constraints, and a concave objective function. The first-order conditions of this problem are therefore sufficient conditions for an equilibrium. If a vector (Q_1, Q_2, r_1, r_2) satisfies the first-order conditions, then it is a global maximum of the problem. There is, however, no guarantee yet that such a point exists, because the constraint set is unbounded - as local production and consumption can be increased simultaneously without violating any constraint. However, the monopolist will never sell an amount that will drive prices down below c_L. A global optimum thus exists, given the assumption that $\lim_{Q\to\infty} P_i(Q) < c_L$.

Optimal Prices The first-order conditions of the optimization problem determine the prices p_i^{NA} that the monopolist sets in region i. In the low-cost region 2, the price is always equal to the local monopoly price $p_2^{NA} = p_{2L}^M$. In region 1, the price p_1^{NA} depends on the size of transmission capacity k.

$$p_1^{NA} = \begin{cases} p_{1L}^M & \text{if} \quad p_1^k < p_1^M \\ p_1^k & \text{if} \quad p_{1L}^M \leq p_1^k \leq p_1^M \\ p_{1H}^M & \text{if} \quad p_1^M < p_1^k \end{cases}$$

where p_{1j}^M is defined by equation (32) and $p_1^k = P_1(k)$.

A.3 Optimization program with arbitrage

This subsection describes the market when there is interregional arbitrage (A). In the main text we assumed that the transmission line was congested, and that electricity is always transported from the low- to the high-cost region. Here we drop these assumptions and allow the line not to be congested, and even allow that the monopolist transports energy from the high-cost region to the low-cost region. Such transports will be represented by a negative x. The monopolist solves the following optimization program:

$$\begin{aligned}
\max_{p_1,p_2,x} \pi^A &= (p_1 - c_H)q_1(p_1) + (p_2 - c_L)q_2(p_2) + x(\Delta c - \Delta p) \\
&s.t. \qquad q_1(p_1) \geq \max(0, x) \\
&\qquad\qquad q_2(p_2) \geq \max(0, -x) \\
&\qquad\qquad \begin{cases} \Delta p \geq 0 & \text{if} \quad x = k \\ \Delta p = 0 & \text{if} \quad -k < x < k \\ \Delta p \leq 0 & \text{if} \quad x = -k. \end{cases}
\end{aligned}$$

The monopolist sets the price in each region, and determines (indirectly) the level of transport from region 2 to region 1. Under arbitrage, the use of the transmission line is determined by the action of the arbitrageurs. The last three lines describe their actions, which depend on the regional price difference $\Delta p = p_1 - p_2$. As for the case without arbitrage, the production and consumption levels should be positive in each region.

Since the problem of the monopolist is a non-convex optimization problem, solutions are not straightforward. The objective function is not concave (1) given the product term $x\Delta p$ in the objective function, (2) for low prices in region p_1 (below c_H), the objective function is not concave in prices and (3) the behavior of the arbitrageurs is described by a highly non-convex set of equations.

A.3.1 Splitting the optimization problem of the monopolist

The optimization problem of the monopolist is called a Mathematical Program with Equilibrium Constraints (MPEC, Luo *et al.,* 1996). This type of model is difficult to solve, as the conditions for the transmission market equilibrium are highly non-convex. We will solve the problem by splitting the non-convex feasible set into three price regions. Once the optimization problem has been solved for each of the regions, we then look for a global optimum. We rewrite the optimization problem as follows:

$$\max_{p_1,p_2,x} \quad \pi^A(p_1,p_2,x) \tag{34}$$

$$s.t. \quad (p_1,p_2,x) \in S(p_1,p_2) \tag{35}$$

with $S(p_1,p_2)$ denoting the feasible set of prices and transmission levels:

$$S = \begin{cases} S^{>} & \text{if} \quad p_1 > p_2 \quad \text{price region I} \\ S^{=} & \text{if} \quad p_1 = p_2 \quad \text{price region II} \\ S^{<} & \text{if} \quad p_1 < p_2 \quad \text{price region III} \end{cases} \tag{36}$$

where

$$S^{>} = \{(p_1,p_2,x) \mid q_1(p_1) \geq k, \; q_2(p_2) \geq 0, \text{ and } x = k\}, \tag{37}$$

$$S^{=} = \left\{ (p_1,p_2,x) \;\middle|\; \begin{array}{c} q_1(p_1) \geq \max\{x,0\}, \; q_2(p_2) \geq \max\{-x,0\} \\ -k \leq x \leq k, \; p_1 = p_2 \end{array} \right\} \text{and} \tag{38}$$

$$S^{<} = \{(p_1,p_2,x) \,|q_1(p_1) \geq 0, \; q_2(p_2) \geq k, \text{ and } x = -k\}. \tag{39}$$

A.3.2 Profit-maximizing prices in the three price regions

We now solve the optimization problem of the monopolist for each of the three price regions, and determine whether the first-order conditions are necessary and sufficient conditions for a global optimum in each of these regions.

Subregion 1: $S^{>}$ The monopolist's problem can be transformed with regard to quantities and becomes:

$$\max_{Q_1,Q_2} \pi^A = (P_1(Q_1) - c_H) \cdot (Q_1 - k) + (P_2(Q_2) - c_L) \cdot (Q_2 + k)$$
$$s.t. \quad Q_1 \geq k,\, Q_2 \geq 0.$$

This problem has a concave objective function, and the constraint set is convex. We therefore know that the first-order conditions are sufficient conditions for an equilibrium. Since the feasible set is unbounded, an optimum might not always exist. The monopolist, however, would never sell a quantity that would drive prices down below c_L. A global optimum thus exists, at least as long as $\lim_{Q\to\infty} P(Q) < c_L$. The monopolist will set the prices $p_1^{>}$ and $p_2^{>}$ in regions 1 and 2 that satisfy the Kuhn-Tucker conditions of the optimization problem:

$$p_1^{>} = \min(p_{1H}^{>}, p_1^k)$$
$$p_2^{>} = \min(p_{2L}^{>}, \bar{p}_2)$$

where the prices $p_{1H}^{>}$ and $p_{2L}^{>}$ are defined as $\frac{p_{2L}^{>}-c_L}{p_{2L}^{>}} = \frac{1}{\varepsilon_2}\frac{q_2+k}{q_2}$ and $\frac{p_{1H}^{>}-c_H}{p_{1H}^{>}} = \frac{1}{\varepsilon_1}\frac{q_1-k}{q_1}$. It can easily be shown that the price in region 1, $p_1^{>}$, decreases in k (as both $p_{1H}^{>}$ and p_1^k are decreasing), and that the price $p_2^{>}$ in region 2 is increasing in k. Hence, the price in region 1 is below the autarky price in region 1 ($p_1^{>} < p_1^M$), and the price in region 2 is above the autarky price in region 2 ($p_2^{>} > p_2^M$). Hence,

$$p_1^{>} < p_1^M \text{ and } p_2^{>} > p_2^M. \tag{40}$$

Subregion 3: $S^{<}$ In subregion 3, the monopolist sets a higher price in the low-cost region. As a reaction to these decisions by the monopolist, arbitrageurs will transport energy from the high-cost region to the low-cost region, and x will become negative. The monopolist might have the incentive to set such prices when incentives to price discriminate in favor of the high-cost region are larger than the production-cost incentives.

The optimization problem in this subregion is similar to that in subregion 1. However, as electricity flows in the opposite direction, the variable k enters the formula with the opposite sign. Using a similar argument as before, we can show that the global optimum exists and that the first-order conditions are again necessary and sufficient for this global optimum. The optimal prices in region 2 and 1 are

$$\begin{aligned} p_2^< &= \min(p_2^k, p_{2L}^<) \\ p_1^< &= \min(\bar{p}_1, p_{1H}^<), \end{aligned}$$

where the prices $p_{1H}^<$ and $p_{2L}^<$ are defined as $\frac{p_{2L}^< - c_L}{p_{2L}^<} = \frac{1}{\varepsilon_2}\frac{q_2 - k}{q_2}$ and $\frac{p_{1H}^< - c_H}{p_{1H}^<} = \frac{1}{\varepsilon_1}\frac{q_1 + k}{q_1}$. As before, it can be shown that the price in region 1 $p_1^<$ is increasing in transmission capacity and that the price $p_2^<$ is decreasing in region 2. Hence,

$$p_1^< > p_1^M \text{ and } p_2^< < p_2^M. \tag{41}$$

Subregion 2: $S^=$ In subregion 2, the price difference $\Delta p = p_1 - p_2$ is zero, and the monopolist no longer pays for using the transmission line ($\Delta p x = 0$). Its objective function becomes an increasing function of x, and transport increases production efficiencies. It is obvious that the monopolist will never set $x < 0$, as it would reduce the monopolist's objective and would shrink the feasible set. In the optimum, consumption in region 2 will be positive, as we assume that the reservation price is sufficiently large $\bar{p}_2 \geq p_1^M \geq p_2^M$ to prevent shutdown of the market in region 2. The monopolist's optimization problem simplifies to

$$\begin{aligned}\max_{p,x} \pi^A &= (p - c_H)q_1(p) + (p - c_L)q_2(p) + x\Delta c \\ s.t. \quad & x \leq \min\{q_1, k\}.\end{aligned}$$

The objective function of the monopolist is concave for prices above the marginal cost c_H, but might be non-concave for prices below the marginal cost c_H. Transforming the problem from the price to the quantity domain makes the objective function concave, but then a new non-convex constraint would have to be introduced ($P_1(Q_1) = P_2(Q_2)$). A transformation therefore does not convexify the problem, and several local maxima might exist.

The solution of the problem depends on the size of k, which determines which of the two constraints will be binding in the optimum. For large transmission capacities $k > q_1(p_{totL})$, the transmission constraint is not binding, and the monopolist will set the price

$$p^{=} = p_{totL},$$

where p_{totL} is defined in equation 33. For small transmission capacities ($k < q_1(p_{totL})$), transmission capacity will be fully used ($x = k$) and several local optima might exist. The price is then given by

$$p^{=} = \arg\max_{p \leq p_1^k} q_1(p)(p - c_H) + q_2(p)(p - c_L). \tag{42}$$

A.3.3 Combining the three regions

The monopolist maximizes his profit by comparing the profits it receives in the three price regions. He solves his optimization program with respect to the non-convex constraint set, which consists of a combination of three price regions. In each region there might be a local optimum, and when there is no local optimum, then the supremum is achieved at the price boundary. One of the main assumptions of our paper is that the autarky price is larger in region 1 than it is in region 2 ($p_1^M > p_2^M$). We show now that under this assumption it is possible to derive conditions on the existence of local optima, and on the relative profit that the monopolist achieves in these local optima.

Proposition 7 *If a local optimum exists in price region I ($p_1^> > p_2^>$), then it is chosen by the monopolist. If a local optimum does not exist in this region ($p_1^> < p_2^>$), then the monopolist sets a uniform price $p^=$ as determined by equation 42.*

Proof. The proof consists of five steps. First, we show that there is no local optimum in price region III, and that we therefore only need to compare the local optima in price regions I and II. Second, if $p_1^> < p_2^>$, then the monopolist will set a uniform price ($p_1^=, p_2^=, x^=$). If $p_1^> > p_2^>$ then a local optimum exists in price region I, and the monopolist will compare its profit in this local optimum with the local optimum in price region II. The following steps of the proof will show that when $p_1^> > p_2^>$, then this is the global optimum for the monopolist. In the third step, we show that for a positive price difference the transmission capacity has to be small. Fourth, for such small transmission capacities, a monopolist that uses uniform prices will always congest the line. Fifth, uniform pricing and congesting the line is dominated by setting a positive price difference with congested lines.

(1) The autarky price in region 1 is larger than in region 2. According to equation 41, this implies that $p_1^< > p_2^<$ and suggests, therefore, that there is no local optimum in price region III (by definition).

(2) If $p_1^> < p_2^>$, then a local optimum does not exist in price region I (by definition), and the monopolist will set a uniform price ($p_1^=, p_2^=, x^=$).

(3) Define k^* and p_2^* by the following two equations:

$$\begin{aligned} q_1(p_2^*) &= k^* \\ -q_2'(p_2^*)\quad (p_2^* - c_L) - q_2(p_2^*) &= k^* \end{aligned}$$

If the price difference is positive in price region 1, $p_1^> - p_2^> > 0$, then $k < k^*$, as the opposite assumption would lead to the following contradiction:

$$\forall k > k^* : \quad p_2^> = p_{2L}^> > p_2^* > p_{1k} \geq p_1^> .$$

The equality follows from the fact that for sufficiently large reservation prices, $p_2^> = \min(p_{2L}^>, \bar{p}_2)$. The first inequality follows from a comparison of the definitions for $p_{2L}^>$ and p_2^*. The second inequality follows from the fact that $k > k^*$, and that $p_1(\cdot)$ is downward sloping. The last inequality is determined by the definition of $p_1^> = \min\{p_1^k, p_{1H}^>\}$.

(4) Above, we showed that for small transmission capacity ($k < k^* < q_1(p_{totL})$) the monopolist will congest the line ($x^= = k$) in price region II. See equation 42.

(5) If the line is congested ($x = k$), then the price vector $(p_1^>, p_2^>)$ gives the monopolist a higher profit than any other price combination (p_1, p_2). This follows from the definition of $(p_1^>, p_2^>)$. In step 4 we showed that the monopolist will congest the line when it sets a uniform price ($x^= = k$). This implies that the local profit in price region I is larger than that in price region II, $\pi(p_1^>, p_2^>, x^>) \geq \pi(p_1^=, p_1^=, x^=)$, through revealed preferences by the monopolist. ■

A.4 Summary

This subsection combines the results of the previous subsections, and shows that for sufficiently low transmission capacities, the optimum is given by the first-order conditions used in the main text.

Proposition 8 *For a small transmission capacity k, the monopolist will set the prices described by equations 6, 7, 11 and 12.*

Proof. The previous subsections described the prices the monopolist will set in both regimes. For a small transmission capacity and *no arbitrage*, the monopolist will set the prices $p_2^{NA} = p_{2L}^M$ and $p_1^{NA} = p_{1H}^M$. For this to be the case, demand in region 1 has to be larger than the transmission capacity, so the line is congested (condition 1: $k \leq q_1(p_{1H}^M)$).

For the case with *arbitrage*, prices are given by the equations $p_1^A = p_{1H}^>$ and $p_2^A = p_{2L}^>$ when the price difference between regions 1 and 2 is positive (condition 2: $p_{1H}^> > p_{2L}^>$), the price in region 2 is smaller than the reservation price (condition 3: $p_{2L}^> < \overline{p}_2$), and the demand in region 1 is larger than the transmission capacity (condition 4: $k \leq q_1(p_{1H}^>)$).

If transmission capacity is sufficiently small such that conditions 1 to 4 hold, then the prices given by equations 6, 7, 11 and 12 define the global optimum.

Conditions 3 and 4 are never binding, so one only needs to check conditions 1 and 2 which can be written as follows:

$$\begin{aligned} k &< q_1(p) = -q_1'(p)(p - c_H) \\ k &< -q_2'(p)(p - c_L) - q_2(p) = q_1'(p)(p - c_H) + q_1(p). \end{aligned}$$

It can be shown that there exists a positive k that satisfies these two conditions. ■