DRAFT OF NOVEMBER 21, 2002

Very Preliminary and Incomplete.

TRANSMISSION INVESTMENT:
ALTERNATIVE INSTITUTIONAL FRAMEWORKS

Paul Joskow[*]

Jean Tirole[**]

1. Introduction

The economic analysis of restructured competitive wholesale electricity markets has mostly focused on the organization and functioning of spot markets for energy and other generation services (e.g. operating reserves and frequency regulation). This theoretical and empirical research has focused on, among other things, the organization of day-ahead and real time (balancing) energy markets and associated auction rules, the role of bilateral contracts, congestion management, nodal pricing, physical and financial transmission contracts, and associated market power issues. This work typically takes the transmission network as given, assumes that there is a fixed non-stochastic amount of transmission capacity available on the network, that the available capacity is unaffected by decisions made by the transmission owner and system operator, and that this capacity is common knowledge to all market participants, transmission owners and the system operator. In reality, even in the short run, the capacity of a transmission network is stochastic as a consequence of facility outages and variations in external conditions such as weather. Moreover, the actual capacity of the transmission network under any

[*] MIT.
[**] IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT.

particular set of supply and demand conditions depends on decisions made by the transmission owner (TO) (e.g. maintenance) and the system operator (SO) (e.g. actions designed to achieve target risks of system failures), which may (as in England and Wales) or may not (as in California and PJM) be the same entity.

In the medium and long term as demand grows and new generating capacity is added to replace older less efficient capacity or to meet growing demand efficiently, investments in transmission capacity are likely to be necessary to minimize the overall costs of wholesale electricity supplies, to maintain reliability, to mitigate locational market power, and to improve the performance of competitive wholesale and retail markets. Indeed, most new investments in generation of any significant size must be accompanied by expansions of the transmission network. This paper investigates the strengths and weaknesses of alternative institutional frameworks to govern transmission investments in an industry where generation and transmission are owned and operated by independent entities. We identify numerous likely performance problems that would result from a framework that relies entirely on *merchant transmission* investment financed entirely by congestion charges. We explore performance issues associated with a framework that relies on a *regulated monopoly transmission provider* governed by alternative incentive regulation mechanisms. Finally, we examine whether and how merchant and regulated transmission models can be integrated in a complementary fashion to improve performance.

2. Transmission Investment: Overview

Decisions regarding investments in new generation (including location) and transmission facilities are inherently interdependent. A new generator requires at least some supporting investment to connect it to the network. More interestingly, additional investments to expand network capacity may be inefficient if the increased power flows from the new generator increase network congestion costs, constrain the operation of low-cost generating plants at particular locations, or reduce reliability. In addition, the locations chosen by new generators will depend, in part, on forecasts of network congestion that may affect prices for generation service at different locations over many years into the future.[1] Finally, when there is congestion, market power may be enhanced at particular locations where competition is limited by import constraints into the area. Locational market power leads to inefficiencies from dead-weight losses resulting from deviations of prices from marginal costs, from inefficient entry and other rent-seeking behavior, and from alternative imperfect market power mitigation mechanisms such as price caps.

When the electric power industry was made up of regulated vertically integrated monopolies, decisions about investments in generation and transmission and associated locational decisions were typically made jointly by the same firm, arguably internalising these interdependencies. In addition, potential market power problems that can arise when the prices generators can charge for power are deregulated, were not an issue for regulated vertically

[1] Generator location decisions depend on many variables include the availability and price of land, the availability of cooling water, the costs of transporting fuel, the costs of connecting to the network, and the costs of congestion on the network at different locations. Generators' decisions to continue operating once investments have been sunk are likely to be more sensitive to locational prices for energy and operating reserves.

integrated firms and their investment decisions did not take market power considerations into account (Joskow 2002). Accordingly, in restructured electricity sectors where generation and transmission investment decisions are made independently and power prices deregulated, some governance framework must be found to facilitate efficient coordination of generation and transmission investments and to account for the short run and long run social costs of congestion, changes in reliability and market power.

Despite the importance of developing such a governance structure, and growing problems associated with stimulating transmission investment in many restructured electricity markets, there has been surprisingly little research on the institutions governing transmission investment in restructured competitive wholesale electricity markets. Early formulations of the structure for competitive wholesale markets envisioned the creation of independent regulated regional transmission and system operating entities (Transcos) that would be responsible for building, owning and operating transmission facilities and would be subject to economic regulation (Joskow and Schmalensee, 1983). More recent research has explored the attributes of incentive regulatory mechanisms that could be applied to such regulated transmission monopolies (e.g. Celebi, Nasser 1997, Léautier 2000, Vogelsang 2001) to integrate energy price (congestion) signals with transmission investment. We refer to this approach as a *regulated Transco* (or regulated Transmission Company) model. The institutional arrangements governing transmission operation and investment in England and Wales reflect this basic institutional approach.

The key issue with the regulated Transco model is then to provide the Transco with incentives to properly maintain and enhance the capacity of the network to deliver power from points of injection to points of consumption. One possibility is a bonified cost-of-service arrangement in which the Transco's costs are reimbursed through charges allocated between consumers at the demand nodes and generators at the supply nodes in a way that minimizes behavioral distortions, subject, perhaps to the provision of specific incentives (for example, to minimize the time it takes to repair a line after an outage). This cost-of-service mechanism may be complemented, to determine new investment, by a stakeholder process, in which the generators and load serving entities express opinions and perhaps even contribute to the build-up of further capacity. Alternatively, the Transco may be given more freedom and more incentives, and, within certain bounds, set the new investments and maintenance policies itself. As with all problems in regulation, the fundamental problems of regulatory mechanism design are associated with asymmetric information between the regulator and the transmission monopoly and associated moral hazard and adverse selection problems (performance incentives and rent extraction).

As an alternative (or complement) to the regulated Transco model, decentralized property-rights based institutions have been proposed to govern transmission investment (Hogan 1992; Bushnell and Stoft 1996,1997; Chao and Peck 1996). These approaches envision new transmission investment creating transmission rights (either physical or financial as described in Joskow and Tirole 2000) based on the capacity of the network to transfer power from points of injection to points of consumption. An investment that increases network capacity would be rewarded with the associated incremental transmission rights. The value of these transmission rights, which are typically equated to the expected congestion charges either avoided (physical rights) or rebated by the system operator (financial rights) over the life of the transmission investment, then provides the financial incentive to invest in new transmission capacity. We will call this the *merchant transmission* model.

Research on this model has focused almost entirely on simple cases where transmission investments are characterized by no increasing returns to scale, there are no sunk cost or asset specificity issues, nodal energy prices fully reflect consumers' willingness to pay for energy and reliability, all network externalities are internalised in nodal prices, there is no uncertainty over congestion rents, there is no market power, markets are always cleared by prices, there is a full set of futures markets, and the TO/SO has no discretion to affect the effective transmission capacity and nodal prices over time. That is, the analysis has proceeded under assumptions equivalent to those of a simple model of perfect competition. Under these assumptions it can be demonstrated (a) that efficient transmission investments that create transmission rights satisfying certain simultaneous feasibility constraints will be profitable and (b) that inefficient transmission investments will not be profitable (Hogan 1992; Bushnell and Stoft 1996,1997). These two results are the primary economic foundation for relying on a merchant transmission model. While there has been some recognition that relaxing these assumptions undermines key results regarding the optimality of merchant investment (e.g. Bushnell and Stoft 1996, 1997; Oren et al 1995), little analysis of more realistic cases has been forthcoming (Perez-Arriaga et al., 1995 is an exception).

No restructured electric power industry has adopted a pure merchant transmission model of this type, though Australia has adopted a mixed merchant and regulated transmission model.[2] However, recent academic proposals[3], as well as FERC's July 2002 Standard Market Design (SMD) proposals, call for relying primarily on "market driven" transmission investments, while recognizing that at least some regulated transmission investments may be necessary. The extreme version of market driven investment allows for free entry into the activity of

[2] Two merchant lines supported by differences in spot prices in the two market areas they connect have been placed in operation under this arrangement in Australia. Directlink is a 180 Mw, 40 mile merchant DC link connecting Queensland and New South Wales and began operating in 2000. Murraylink is a 220 Mw, 108 mile merchant DC link connecting South Australia and Victoria which began operating in October 2002. On October 18, 2002, Murraylink applied to the regulatory authorities in Australia to change its status from a merchant line to a regulated line that would be compensated based on traditional cost of service principles combined with a performance incentive mechanism. Neither merchant link appears to be profitable. As far as we can tell, these are the only two merchant transmission lines that have been built in anticipation of recovering their costs entirely from congestion rents arising from the difference in nodal prices operating anywhere in the world.

[3] e.g., Hogan (2002).

constructing transmission lines. The owners of these transmission lines are rewarded through the congestion rents attached to the lines.

Transmission investment institutions cannot be considered independently of the institutions that govern the determination of energy prices, operating reserves, contingency constraints, congestion management, and the specification of transmission capacity and increments to it. No single paradigm has emerged from the liberalization effort of the last decade for these attributes of the operation of wholesale markets, system operations, and congestion management. So, we need to be more precise about the organization of the wholesale market, congestion management and price determination to understand and evaluate alternative institutional frameworks to govern transmission investment. In what follows we will assume that a nodal pricing system is in place with attributes similar to those being proposed by the U.S. Federal Energy Regulatory Commission (FERC) in its SMD proposals and what is in operation in New York and PJM in the U.S. This is the most conducive framework for merchant investment because nodal prices provide a measure of locational scarcity.

Under this model, an independent system operator (ISO)[4] operates a real-time balancing market and manages congestion. The ISO takes all of the bids (generation and demand) and finds the "least cost" set of uniform market-clearing price bids to balance supply and demand at each generation and consumption node on the network using a security constrained dispatch model. This establishes day-ahead quantity commitments and nodal prices with deviations from day-ahead schedules settled in similarly structured real time or balancing markets. The resulting nodal prices reflect both congestion and marginal losses. Generators may enter into bilateral contracts with marketers or load serving entities (LSEs) and schedule supplies with the ISO

[4] Renamed Independent Transmission Provider (ITP) in the FERC proposal.

separately from the organized day-ahead market. However, they still have to pay any congestion charges associated with their schedules based on the difference in nodal prices between the injection and receipt points. The day-ahead schedules, nodal prices, and congestion charges are "commitments." They can be adjusted in real time by submitting adjustment bids to the real time balancing markets (which again rely on bids, a security constrained dispatch and nodal prices) to allow these schedules to be changed based on real time economic conditions. The model recognizes that there are incumbent transmission owners (TO) that own the existing transmission assets and requires that the SO and TO be separate entities and operate independently. The TO receives some cost-of-service compensation for the usage of a grid that it no longer controls to compensate it for legacy investments and ongoing maintenance costs. New investments in transmission are anticipated to be made by competing merchant investors whose compensation is based on the value of Congestion Revenue Rights (CRRs)[5] created by their investments. These financial rights represent the right to receive congestion revenues defined as the difference between the nodal prices between the two nodes (point-to-point) covered by the relevant CRR times the quantity of CRRs held. In Joskow and Tirole (2000) we defined these rights as representing a *share* of the congestion revenues (or merchandizing surplus) earned by the system operator. This formulation implies that the obligations to pay rights holders is always the same as the congestion revenues earned by the system operator. Under the CRR formulation, however, the quantity of point to point financial rights is fixed ex ante and allocated to holders to reflect estimates of the capacity of the network to accommodate schedules that fully utilize these rights. In this case, deviations between actual transmission capacity and the number of allocated rights results in the congestion revenues earned by the SO being either too little to fully cover the

[5] Congestion Revenue Rights is the name FERC has now given to financial rights that have been referred to in the literature as Transmission Congestion Contracts (TCCs) or Financial Transmission Rights (FTRs).

associated financial obligations to rights holders or to congestion revenues in excess of what is owed to rights holders. For example, if K rights are issued to inject power at node 1 and receive it at node 2, the rights holders are owed $K(p_1 – p_2)$, where p_1 and p_2 are the prices at the 2 nodes. If the actual capacity of the network turns out to be K_a then the system operator will have a congestion revenue or surplus equal to $(K – K_a) (p_1 – p_2)$.

In the U.S. and Australia, proposals for market driven transmission investment are supplemented with options for incumbent TOs to make regulated transmission investmemts as well. This raises difficult issues associated with efficiently mixing market driven and regulated transmission investments on the same network.[6] In Australia, this mixture of competition and regulation has led to extensive litigation between proponents of regulated and merchant transmission links, delaying investments in both.

The separation between ownership (TO) and control (system operating or SO) functions in this model is motivated by two considerations. First, a market driven transmission system leads to multiple owners of parts of the grid; while the owners can form a cooperative to operate the grid, their goals are in general antagonistic,[7] and it is well-known that cooperatives of members with heterogeneous interests face complex governance problems.[8] Second, and quite crucially, grid owners face a serous potential conflict of interest when operating a transmission grid *if* their compensation varies directly with the level of congestion rents. In practice, due to the lack of market-based penalties for outages, dispatching does not quite correspond to the least-cost optimization used in economic and engineering models; rather, grid operators have

[6] The possibility that a merchant link may be able to switch to regulated cost-of-service status ex post, as is proposed by MurrayLink in Australia, also raises interesting incentive issues.

[7] Incumbent owners of transmission lines that are being compensated based on congestion rents will have incentives to oppose investments by others in generation or transmission that reduce these congestion rents. Generators located in congested areas will have incentives to oppose transmission enhancements that would reduce or eliminate the congestion.

[8] See, e.g., Hansmann (1996).

substantial discretion over how much outage they are willing to take while dispatching.[9] This discretion in turn potentially provides incentives for system operators to manipulate the congestion rents received by the owners (Glachant and Pignon, 2002). By conservatively "withdrawing" transmission capacity (under the cover of a safe management of the network), the system operator can substantially raise the congestion rents. Finally, in many countries there continues to be vertical integration between generation and transmission. The creation of an independent SO is thought to be a way to mitigate the potential problems that may arise from common ownership and control of transmission and generating assets and associated power marketing activities. However, the separation of ownership (TO) and operations (SO) carries other potential costs caused by inefficient coordination between the SO and the TO. Accordingly, there is a tradeoff between integration of TO/SO functions and separation of these functions that has largely been ignored in the literature and by policymakers. We examine the resulting “moral hazard in teams” issues further below.

Finally, the restructuring of electric power systems to rely on competitive wholesale markets does not start with a blank slate. There generally exists an extensive legacy transmission network and an associated fleet of generating plants. The configuration of these assets may not be “optimal” in the ex ante sense for at least two reasons. First, supply and demand conditions are likely to have changed from what was assumed when these investments were made. Second, the investments were not made to be optimally configured to accommodate a decentralized competitive wholesale market. For example, vertically integrated firms would not have taken local market power problems into account since they would have had no incentive to exercise market power against themselves (Joskow 2002). The configuration of the legacy network must

[9] For example, the so-called (N-1) and (N-2) constraints are self-imposed constraints and are subjective responses to the perceived risk.

be taken into account in the evaluation of alternative institutional arrangements to govern its operation and investments to expand its capabilities.

For these reasons, we have found it useful to consider two types of transmission investments that can increase the capacity of the network (or, alternatively reduce congestion) to accept injections of energy at a particular location A on the network for consumption at another point B on the network.[10]

Network deepening investments: These are investments that involve physical upgrades of the facilities on the incumbent's existing network (e.g. adding capacitor banks, phase shifters, reconductoring existing transmission links, new communications and relay equipment spread around the network to increase the speed with which the SO can respond to sudden equipment outages and relax contingency constraints). These are investments that are physically intertwined with the incumbent TO's facilities. These investments are specific investments (as described by Williamson 1983) that we assume can be undertaken most efficiently by the incumbent network owner. Similar to network deepening investments are *network maintenance* decisions. Like network deepening, maintenance is most efficiently performed by the owner of the link or grid. We will discuss "access pricing" and related arrangements that might be used to allow for competitive provision of network deepening investment and maintenance further below.

[10] We focus on transmission investments that affect congestion on the high voltage network.Regulators in the U.S. often break transmission investments down into additional categories. First, there are local transmission investments "inside" the demand node. These investments are sometimes called "reliability" investments. Second, interconnection investments are investments that must be made by an incumbent grid owner to connect new generators with the rest of the network. These are often treated like radial links and are typically paid for by the generators seeking interconnection. However, it is hard to draw bright lines between reliability investments, interconnection investments, and investments that affect network congestion.

Independent network expansion investments: These are investments that involve the construction of separate new links (including parallel links) that are not physically intertwined with the incumbent network except at the point at either end where they are interconnected. These investments can (in principle) be made either by incumbent transmission owners, by stakeholders (generators, load-serving entities), or by a third-party merchant investor. The two operating DC merchant links in Australia appear to fall into this category. However, as in Australia, these links may have effects on power flows on the rest of the network, including on parallel lines, but are physically separable projects from a construction and maintenance perspective.

In addition to legacy transmission assets there are also typically legacy obligations and rights that accrue to various market participants including consumers in conjunction with the restructuring process and transition to reliance on competitive wholesale and retail markets. For example, the incumbent owners of the transmission network built under regulation expect to be compensated for the regulatory asset value (embedded costs) of the associated facilities. Consumers may have certain legacy rights to receive power and predetermined prices and load serving entities (e.g. distribution companies) obligations to provide these services. These rights and obligations may include “firm rights” to use the legacy transmission assets free from congestion charges. These legacy rights and obligations place constraints on transmission prices and the allocation of transmission rights and can have an important effect on the implementation of alternative models for transmission investment.

3. Market-driven or “Merchant” transmission investment

Let us start from the theoretical case for market driven or “merchant” transmission investment (this rationale has been developed, inter alia, by Hogan 1992 and Wu et al 1996, Chao and Peck 1996 and examined further by Bushnell and Stoft 1996, 1997 for simple cases.) The basic argument is conveyed in the two-node framework of figure 1.

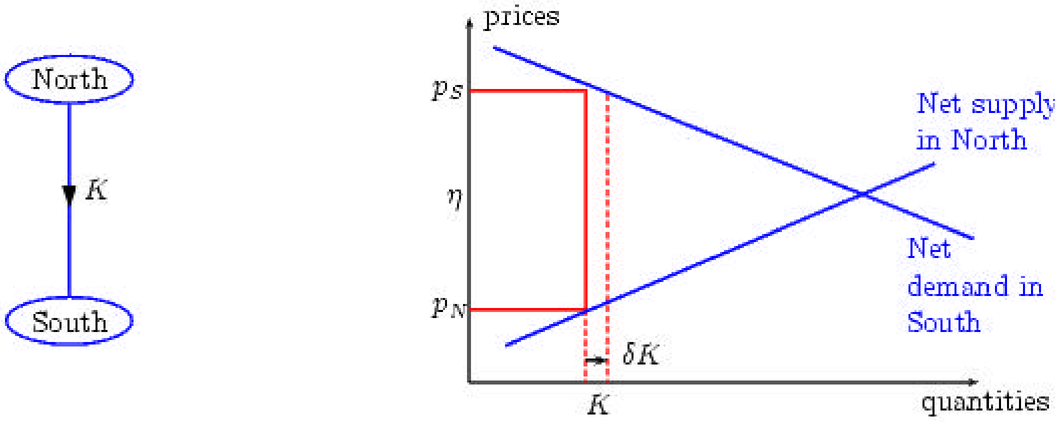

Figure 1

Figure 1 depicts a simple situation in which industrial users and load serving entities in the South (say, a large city) buy their power from cheap generation sources in the North and, possibly, more expensive sources in the South. Alas, the capacity of the line from North to South is limited to *K*, and faced with net demand /supply curves in the North and the South, the system operator is forced to dispatch "out of merit". For example, the system operator calls on expansive generators in the South while generators in the North would be willing to supply this amount at a lower price. The rationing of the scarce North-South capacity is implemented by setting two nodal prices, p_S and p_N that clear the markets in the South and the North, respectively. The difference, $h = p_S - p_N$, is the shadow price of the transmission capacity constraint (or the congestion rent).

Now consider a *marginal* (unit) increase in transmission capacity. This unit increase allows one more KWh to flow from North to South, replacing a marginal generator in the South with cost p_S by a cheaper generator in the North producing at cost p_N. Assume that the builder of this marginal capacity, whether it is a new entrant or the incumbent TO, is rewarded through a financial transmission right that pays a dividend equal to the shadow price of the transmission constraint. An independent merchant company will build this extra capacity as long as h exceeds the cost of building it. By contrast, an incumbent grid owner that is compensated through the payment of congestion rents, may not want to make this marginal investment as it must compare the extra revenue h net of the cost of expanding the capacity with the reduction in the congestion rent on its inframarginal transmission units $(-Kdh/dK)$.[11] It is only when the incumbent grid owner's capacity has been rated at some level K^* not too different from actual capacity, and that the corresponding rights, with value hK^*, have been auctioned off, that the monopoly distortion vanishes. The incremental capacity then yields $h + (K - K^*)\frac{dh}{dK}$, close to h. As in the case of Contracts for Differences,[12] forward sales restore proper incentives for a player with market power.

Hogan (1992) and Bushnell and Stoft (1996, 1997) show that under certain conditions (e.g. no increasing returns to scale, simultaneous feasibility constraints bind when awarding congestion rights, efficient nodal prices clear all markets, no market power in the wholesale

[11] Of course, the regulated incumbent grid owner could be rewarded in a way that ties its reward to *reducing* congestion rather than increasing it. We will discuss this type of regulatory mechanism below. It should be recognized, however, that once a merchant project is built it becomes an incumbent that will find further investments that reduce congestion rents to be unattractive.

market, well defined property rights, a complete set of competitive liquid forward markets to provide sufficient statistics for long run demand and supply conditions and risk management, etc.) all *efficient* transmission investments will at least recover their costs from congestion revenues and that *inefficient* investments will not be profitable. These are potentially powerful results that appear to transform the transmission investment problem from one that appears to be almost intractable to one that requires a simple implementation of a property-rights based market system.

Merchant investment's appeal is that it allows unfettered competition to invest in new transmission capacity, placing the risks of investment inefficiencies and cost overruns on investors rather than consumers, and bypassing planning and regulatory issues associated with a structure that relies on regulated monopoly transmission companies. In addition, in theory, it allows investment in new generating capacity in the constrained area to “compete” with new transmission investment that reduces the import constraint. In this way, market driven transmission investment is an economist’s dream, solving the problems associated with imperfect regulation of a “natural monopoly” transmission company and aligning competitive transmission investments with the newly developed competition in the generation segment. Unfortunately, the optimality of the market driven approach depends on a number of strong assumptions and conditions that are likely to be inconsistent with the actual attributes of transmission investments and the operation of wholesale markets in practice. (some of the critiques will apply to alternative frameworks as well). We turn now to a discussion of what we view as the most important assumptions underlying the case for the merchant model and the implications of relaxing these assumptions. We will assume that wholesale markets are organized around the “nodal pricing” model utilized in PJM, New York and proposed in FERC’s SMD. However, we

[12] See Green (1992).

will ignore issues associated with common ownership of generation and transmission by assuming that the TO and SO are independent. However, unlike much of the analysis underlying this model we will recognize that the TO and SO have objective functions, can make discretionary decisions that can affect market performance, including reliability risks, respond to incentives they face (including political pressures in the case of non-profit SOs), and that TO and SO decisions may be interdependent leading to potential costs of imperfect coordination between them.

a) *Competitive energy markets that clear with efficient prices*

The reasoning above assumes that the prices that clear the markets in the North and the South reflect the marginal costs of production (and the marginal willingnesses to pay[13]), so that the congestion rent perceived by merchant investors does reflect the social savings brought about by the investment. That is, potential investors in new transmission capacity see the correct locational price signals in the wholesale markets. There are a number of reasons why this is unlikely to be the case.

Suppose for example that there is a generator with market power in the South, and that the latter region is import constrained. The generator exercises market power by withdrawing capacity and driving the price in the South up. Hence

$$p_S > \mathrm{c}_S \ ,$$

where c_S denotes the marginal cost of production in the South. The measured congestion rent then overestimates the cost savings associated with the replacement of one unit of power generated in the South by one unit of power generated in the North, suggesting an *over-incentive*

[13] We will present the argument in terms of cost savings; because what matters is net supply at each node, the same argument would apply to the demand side.

to reinforce the link. On the other hand, the same inequality suggests that the additional power flowing through the link may result in a “business stealing effect,” whereby it crowds out output yielding a margin to the monopolist in the South. The box below shows that, under a very weak assumption,[14] the first effect in general dominates, and therefore market power results in an over-incentive to invest.[Similarly, and to the extent that reinforcing the line is akin to adding production capacity in the South, this suggests that entrants in generation have too much of an incentive to invest in the South. The box verifies that this is indeed the case]. Thus, despite market power in the load pocket (which suggests that relieving congestion through expansion in transmission is a public good), market signals provide no under-investment incentives.

The impact of locational market power on merchant investment incentives

• Consider a monopoly supplier in the South producing at marginal cost c_S and facing demand function $D(p_S)$.

Provided that the transmission capacy K between North and South is fully utilized, the monopolist solves :

$$\max_{p_s}\left\{(p_S - c_S)\left[D(p_S) - K\right]\right\};$$

equivalently, this monopolist selects a consumption q_S in the South so as to solve :

$$\max_{q_s}\left\{\left[P(q_S) - c_S\right](q_S - K)\right\},$$

where $P(.)$ denotes the inverse demand function. Neglecting consumption in the North, social surplus is

$$W = S^g(q_S) - c_N K - c_S[q_S - K].$$

The marginal gross surplus, dS^g/dq_S is equal to price p_S in the South, and so when the line’s capacity is increased by dK, resulting in a consumption change dq_s, welfare changes by

[14] The assumption is that the Southern monopolist's reaction curve be downward sloping in a Cournot game. Intuitively, the transmission line creates a Cournot "duopoly" in the South, in which the Southern firm faces a fixed output from its (transmission) rival. A downward sloping reaction curve means that the Southern firm curtails its output as the transmission capacity expands. This implies that the business stealing effect is smaller than the inflated signal effect (the two effects would cancel if the output in the South were invariant to an increase in imports from the North).

$$dW = (p_S - c_S)dq_S + (c_S - c_N)dK.$$

Note that,with *perfect competition*, $p_S = c_S$ (and $p_N = c_N$) and so $dW = hdK.$

With monopoly power in the South, though,

$p_S - c_S > 0$

and

$c_S - c_N < h.$

There is an over-incentive to invest if and only if

$dW < hdK,$

and

$(p_S - c_S)dq_S + (c_S - c_N)dK < (p_S - c_N)dK,$

that is if and only if

$$dq_S < dK.$$

For there to be an over-incentive to invest, the monopolist must « absorb » some of the increase in inports from the North. To know whether this is the case, differentiate the first-order condition for profit maximization :

$$\frac{dq_S}{dK} = \frac{P'}{2P' + (q_S - K)P''}.$$

Thus, there is an over-incentive to invest under merchant investment if and only if

$$P' + (q_S - K)P'' < 0.$$

A sufficient condition for this is that the demand curve be concave. More generally, this condition is the standard condition for quantities to be strategic substitutes.

- The same reasoning can be applied to generation investments in the South. Indeed, *K* could alternatively denote the amount of power produced by a competitive fringe in the South in the profit maximization exercise. And so

$$dq_S < dK$$

as long as

$$P' + (q_S - K)P'' < 0.$$

There is an over-incentive to invest if and only if

$$dW - (p_S - c_S)\,dK < 0$$

or

$$\begin{aligned} d\left[S^n(q_S) - c_S q_S\right] &< (p_S - c_S)\,dK \\ \Leftrightarrow \quad (p_S - c_S)(dK - dq_S) &> 0. \end{aligned}$$

Hence, there is in general an over-incentive to invest in generation in the South as well.

Conversely, a generator with market power in the North may (while still making full use of the link) be able to raise price p_N by withdrawing production capacity — perhaps to the level of p_S if it faces no competition in the North (Oren, 1997; Stoft, 1999; Joskow and Tirole 2000). In this case, the congestion rent underestimates the gain from expanding the line's capacity, resulting in an *under-investment* by merchant transmission investors. At the same time, it could lead to inefficient entry of generating capacity in the North in response to the short run monopoly rents created there.

Returning to the case of market power in the South (this is the situation that will generate very high prices for consumers in the South), the regulator may be tempted to impose a price cap.[15] While the price cap improves economic efficiency if it really is about constraining market power, it may also distort price signals if high prices are due to poor conditions rather than price manipulation. A cap $p_S \leq \overline{p}_S$ then reduces the congestion rents during those hours that are very important because they produce the bulk of the rents to support investment, yielding *under-investment* in transmission.

[15] Or a de facto price cap as when the system operator curtails load administratively when prices don't clear the market.

The FERC SMD proposes to cap day-ahead and real time prices (the suggested cap is $1000/Mwh) and to impose additional market power mitigation, including bid-restrictions and must-offer obligations, to deal with local market power problems such as those discussed for the North and the South above.[16] The SMD recognizes that these caps may "clip" high prices that properly reflect competitive scarcity values as it endeavours to constrain prices reflecting market power and lead to under-investment in generating capacity. To deal with this problem, the SMD proposes that retail load serving entities (LSEs) be required to take on a capacity obligation (e.g. 115% of peak load) to ensure that there is enough generating (or demand reduction) capability to provide "adequate" levels of reliability. This requirement is accompanied by a proposed enforcement mechanism. If the resource adequacy requirements are binding constraints, there will be a market price for "capacity" that meets the specified capacity resource criteria. The price for this capacity then should serve as a sort of safety valve to provide revenues necessary to attract enough investment to meet the specified reserve margin quantity target. However, the SMD does not recognize directly that these price caps and related market-power mitigation measures will also distort incentives for merchant transmission investment by constraining locational prices. The SMD however does include a "deliverability" requirement for resources satisfying the resource adequacy test (capacity obligations), which may provide an independent incentive for generators to pay for expansion of the transmission network.

Prices may not clear supply and demand in real time because market clearing processes are not fast enough to respond to rapid changes in supply and demand conditions while

[16] The FERC SMD would require that under certain "non-competitive conditions" (e.g. local market power problems caused by congestion) generators be required to offer all available energy (must-offer requirement) to the system operator subject to a pre-specified bid cap. FERC Docket No. RM01-12-000, Notice of Proposed Rulemaking, July 31, 2002, paragraph 409. It also invites ITPs to propose additional mitigation measures that could apply under certain conditions where market power would be a significant problem, id. at paragraph 415. Finally, the SMD provides for a regional "safety net bid cap" that would apply to the day-ahead and real time markets under all conditions, id. at paragraph 433.

maintaining physical requirements for frequency, voltage, and stability on the network. To maintain physical network parameters, administrative rationing is then substituted for prices to balance supply and demand as a consequence of what is effectively a problem of incomplete markets (Wilson 2002). Whether it is administrative rationing in response to incomplete markets or price controls motivated by efforts to constraint market power or price distortions caused by market power or discretionary decisions by the system operator, actual prices will depart from the efficient prices required to give the efficient signals for new investment. These imperfections are potentially important with regard to transmission (and generation) investment because the prices that create significant congestion rents tend to occur in a relatively small number of hours and these hours also happen to be the hours when these types of price distortions are most likely to occur.

b) *No lumpiness*

Network expansion investments are likely to be lumpy. That is, the average cost of a new link declines as its capacity increases, other things equal (Baldick and Kahn 1992, Perez-Arriaga, et. al. 1999). [Many *network deepening investments* are less lumpy, but these investments are most conducive to investment by the incumbent network owner rather than a merchant entrant. We discuss deepening investments further below.] The impact of lumpiness is illustrated in figure 2. The initial capacity is K_0 and economically can be brought to a level K_1.

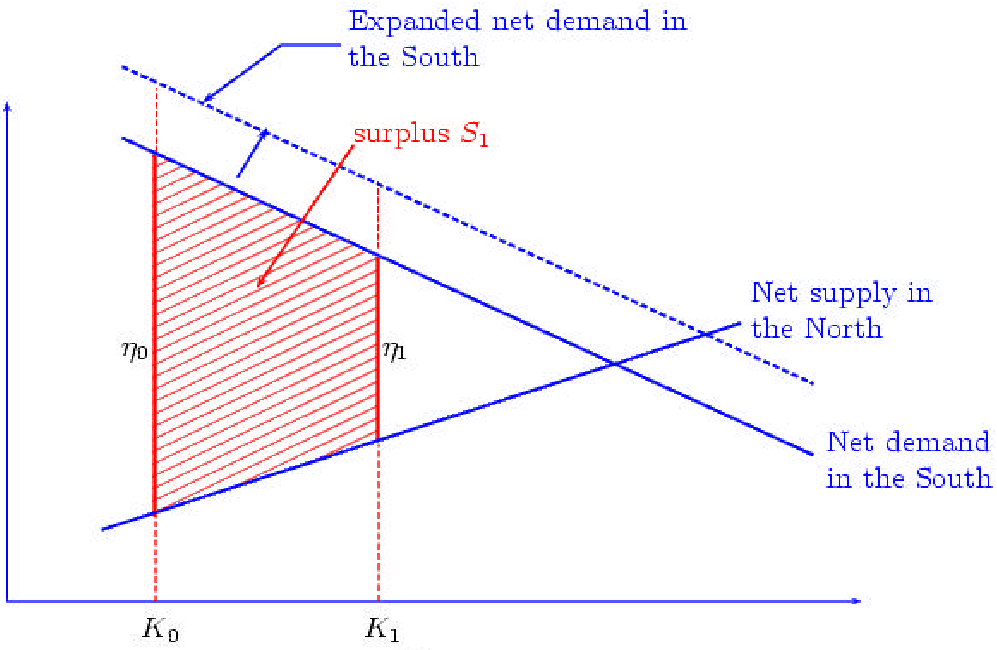

Figure 2

Assuming that the energy market participants are perfectly competitive, so that net demand/supply curves represent the true marginal costs/willingnesses to pay, and the market is cleared by efficient nodal prices, the surplus created by the expansion of capacity from K_0 to K_1 is depicted by the shaded area in figure 2. The value, $h_1(K_1 - K_0)$, of the transmission rights granted to the merchant investors building this capacity expansion understate the social surplus it creates. Lumpiness thus results in an *underincentive* to reinforce the system for the same reason that an incumbent grid owner rewarded by congestion rents has suboptimal incentives to remove these congestion rents.

Another source of lumpiness for *network expansion* investments arises because there may be a *scarcity of rights of way*, for example a unique corridor between a cheap and an expensive area. The difficulties that new transmission corridors face in obtaining siting authority suggests that the available corridors for new lines through many areas will be limited in the sense that (for example) one additional corridor may be available through the Pyrenees, and it may accommodate *one* new link that could be of any size between 100 MW and 1000 MW. This scarcity is particularly problematic as demand grows. Merchant investment is then likely to end up in a "preemption and monopoly" situation. A merchant will install a small capacity on the corridor and will later expand this capacity (presumably, the merchant will underinvest in this expansion as we have seen), to the extent that the expansions are now deepening investments. To be certain, under perfect competition, rents will be dissipated through a very early entry into the scarce corridor, or, if the corridor is put up for auction, through high bids. But the outcome is then similar to a monopoly outcome. Moreover, scarce corridors are typically not allocated through auctions but rather through a regulatory process that places a premium on being first in line.

Besides generating too little investment, lumpiness also may make merchant investment occur too early when it takes place in order to pre-empt additional entry. In a system with growing demand, pre-emption leads to an investment at the first date at which the discounted value of the financial rights on the additional capacity is equal to the investment cost. It could also lead to the investment being "undersized". For example, if the optimal investment is 600 MW, a merchant developer may find it most profitable to invest in a 300 MW enhancement, pre-empting additional investment. The box contains an analysis of the incentive to get a toehold by sinking a small investment.

Lumpy investments: preemption and toeholds

Suppose that at some future date *T*, the net demand in the South jumps up to a new level (the dotted line in figure 2); the post-reinforcement shadow price jumps from h_1 to $h_2 > h_1$. Letting *r* denote the interest rate and *I* the investment cost, suppose that

$$h_1\left(K_1 - K_0\right) < rI < h_2\left(K_1 - K_0\right).$$

Then, under free entry into merchant investment, investment occurs at date $t < T$ such that

$$\left[\frac{1-e^{-r(T-t)}}{r}h_1+\frac{e^{-r(T-t)}}{r}h_2\right](K_1-K_0)=I.$$

Note that this preemption is actually socially beneficial if the surplus S_1 brought about by the expansion before the increase in demand exceeds the interest on the investment cost, i.e.:

$$S_1 > rI.$$

Otherwise, preemption is socially wasteful.

And the point about underinvestment remains: Letting S_2 denote the surplus after demand has grown, if

$$h_2(K_1-K_0) < rI < S_2,$$

then no merchant investment ever takes place even though it is socially desirable.

Similarly, we can show that preemption may encourage inefficiently small investments. Suppose that capacity K_1 can either be reached in one stage, at cost I, as discussed above, or in two stages. The first stage costs I' and yields capacity K', $K_0 < K' < K_1$, which can then be upgraded at cost $I”$ to K_1.

Let $h' \in (h_1, h_2)$ denote the congestion cost for capacity K' *before* demand in the South jumps up. Let us look for an equilibrium in which a merchant investor preempts at date $t < T$ by investing a little (I') and then upgrades the line at time T :

$$\begin{aligned} I' = \quad & \frac{1-e^{-r(T-t)}}{r}h'(K'-K_0) \\ & +e^{-r(T-t)}\left[\frac{h_2(K_1-K_0)}{r} - I''\right] \end{aligned}$$

For this, it must be the case that preemption at $(t-e)$ with the full investment does not pay off:

$$I \geq \left[\frac{1-e^{-r(T-t)}}{r}h_1+\frac{e^{-r(T-t)}}{r}h_2\right](K_1-K_0),$$

or

$$\begin{aligned} & I - \left[I' + e^{-r(T-t)}I''\right] \\ & \quad \geq \left[\frac{1-e^{-r(T-t)}}{r}\right]\left[h_1(K_1-K_0)-h'(K'-K_0)\right] . \end{aligned}$$

Note that the right-hand side of this inequality is negative if the total value of the rights (the total congestion rent) decreases with the capacity of the link..

Aside from the timing considerations discussed above, note that given an entry at t, a social planner might want to jump to capacity K_1 directly, as the social surplus is larger under capacity K_1 than under capacity K'.

c) *Perfect coordination of interdependent investments in generation and transmission*

As noted earlier, the optimality of merchant investment requires that the net demand and supply curves in the wholesale market represent the true demands and supplies of energy market participants. Most of the literature supporting transmission investment is static in the sense (a) there is no uncertainty about supply, demand or prices, and (b) all investments in generation and transmission occur simultaneously. However, investments in transmission are long-lived sunk investments and their value depends on changing and uncertain supply and demand conditions over many future years. The economic calculus necessarily involves forecasting future supply and demand conditions which are uncertain, including changes in locational supply and demand conditions resulting from future investments in generating and transmission capacity, and the associated uncertain nodal prices. As a result, the presentation of the supply and demand functions in the previous figures, and the standard formulations of these problems must stand for the *long-term* demand and supply curves. The latter of course reflect the possibility of investments in generation and (for consumers) bypass. Existing and new electricity producers formulate investment plans, whose implementation depends on the expectations of market conditions; similarly large and small users may adopt equipments that allow them to switch to alternative sources of energy. For example, investments in the North will be unprofitable if they are not accompanied by a strengthening of the North-South line. Conversely, the reinforcement of the line won't be profitable if the congestion rent is too small, that is if no investment occurs in the North.

In principle, this coordination can be achieved through a planning procedure, in which all interested parties announce their (price-contingent) investment plans. Such coordination however becomes more involved if either some party (or coalition of parties) have market power or an incentive to block investments to create it or if investments are lumpy. Mechanisms designed to aggregate stakeholder preferences to make choices about major transmission investments have not been particularly successful.[17]

For example, the owner(s) of the existing capacity K_0 of the line may announce a substantial reinforcement in the hope of attracting investment in the North, and later not implement this capacity building. The price collapse in the North brought about by "excessive" investment in generation there increases the congestion rent and benefits the transmission owner. Similarly, investments in generation in the South might be announced, that are meant to preempt a reinforcement of the transmission line and will never be implemented.

In general, proper incentives must be put in place in order to prevent such manipulations of other parties' investments.

d) *Equal access to investment opportunities*

The free entry merchant investment paradigm requires that there effectively be free entry. One can think of at least two situations in which free entry is not a good assumption.

First, *network deepening* investments can, as a practical matter, only be implemented efficiently by the owner of the existing lines. Defining an efficient "competitive access to deepening investments" policy is likely to be extremely difficult for several reasons. First, adding third-party facilities that are fully integrated with the existing network from a physical and maintenance perspective creates significant incentive problems with decentralized ownership.

[17] See Chisari et. al. (2001) for a discussion of the experience in Argentina.

The problems of defining a good set of rules for investing in and maintaining facilities of this type with decentralized ownership is further exacerbated by the heterogeneous nature of transmission facilities. While it is theoretically possible to devise contractual arrangements that will solve the incentive problems, including opportunistic behavior of one or more parties, investments with these attributes are most likely to be governed efficiently through ownership by a single firm ; second and relatedly, one would need to carefully allocate the new capacity of the line between the initial design and maintenance choice of the original owner and the actions of the renters who make deepening investments. This "moral hazard in teams" problem is a substantial obstacle to the design of an effective third party access policy for this type of transmission investment.

This raises the question of how incumbent transmission owners are to participate in a “market driven” transmission investment framework. On the one hand, precluding them from participating would mean that potentially low-cost network deepening investments will be lost. On the other hand, allowing them to make unregulated merchant investments for network deepening enhancements to which they have unique access would allow them to exercise market power, restrict supplies and capture rents that might otherwise go to consumers under a regulated investment regime. It is natural then to think about allowing incumbents to make regulated investments and new entrants make merchant investments. However, to the extent that the regulated and merchant investments involve parallel lines[18] whether or not the most efficient investments are made will depend heavily on the regulatory mechanisms adopted. Mixing regulated and unregulated activities that are (effectively) in competition with one another is always a very challenging problem.

e) *Objective definition of existing transmission capacity and incremental capacity (1)*

The analysis of merchant investment assumes that the capacities K_0 and K_1 are well-defined and non-stochastic. This abstracts from some important issues that arise even in the two-node case, but are especially problematic in more complex networks with loop flow, which we discuss in a separate section below. In practice, even in the two-node model, the actual capacity of the North/South link depends on exogenous environmental parameters; furthermore, system operators have substantial discretion on defining and implementing security constraints, affecting the actual power flows on the link in real time. For example, the physical capability of transmission lines depends on temperature and other exogenous contingencies.[19] And, of course, even a well-maintained system will have some random outages that cause the available capacity of the link to be reduced.

This all raises the issue of the number of financial rights to be allocated for the existing system and as a consequence of new investments, how congestion revenue deficiencies or surpluses arising from deviations between the number of rights allocated ex ante and the actual capacity of the network ex post are handled,[20] and how these allocation and compensation decisions affect investment and the ultimate performance of the system. We consider this issue

[18] It is exactly this kind of investment that has led to extensive litigation in Australia.

[19] The rated capacity of Path 15, connecting Northern and Southern California falls by about 600 MW as the ambient temperature rises, other things equal. The rated capacity of Path 15 varies by about 1300 MW depending on the availability of various remedial action schemes to respond to transmission and certain generation outages. California ISO, Operating Procedure T-122A, November 6, 2002. It is also important to recognize that in the U.S. and Europe there is not a single SO controlling the network, but multiple SOs controlling independent segments of the network. To maintain reliability and avoid free riding less flexible contingency criteria must be defined than might be the case if there were a single SO operating the network in real time. For example, the simultaneous import transmission capacity into Southern California varies by 700MW depending on the operating status of the three units of a nuclear generating plant in Arizona. California ISO, Operating Procedure T-103, November 6, 2002.

[20] As discussed earlier, the quantity of point to point rights financial rights is fixed ex ante and allocated to holders to reflect estimates of the capacity of the network to accommodate schedules that fully utilize these rights. In this case, deviations between actual transmission capacity and the number of allocated rights results in the congestion revenues earned by the SO being either too little to fully cover the associated financial obligations to rights holders or to congestion revenues in excess of what is owed to rights holders. For example, if K rights are issued to inject power at node 1 and receive it at node 2, the rights holders are owed $K(p_1 – p_2)$, where p_1 and p_2 are the prices at the 2 nodes. If the actual capacity of the network turns out to be K_a then the system operator will have a congestion revenue or surplus equal to $(K – K_a) (p_1 – p_2)$.

in the simple two-node case and explore the issue further when we consider loop flow below. Suppose that K is stochastic: $K = K(q)$, $K'(q) > 0$ and q is distributed between q^- and q^+. Let's say that the line is congested for all values of q, but the value of h will vary with $K(q)$. For which value of q should one compute the number of financial rights? One could be conservative and set the number of financial rights equal to $K\left(q^-\right)$. One would issue $K\left(q^-\right)$ financial rights and owe the holders $hK\left(q^-\right)$ in congestion payments. When the realized q is q^-, one satisfies the feasibility and revenue adequacy condition. But what happens when $q > q^-$? The merchandising surplus will exceed what is owed to the rights holders. What does one do with the excess and how does the distribution affect investment incentives? At the other extreme, one could set the number of financial rights to reflect the maximum capacity $K\left(q^+\right)$. There would be revenue adequacy when $q = q^+$ but not when $q < q^+$, which would be most of the time since the system operator would owe $hK(q^+)$regardless of the actual realization of q. Where does the shortfall come from and how does it affect investment incentives? The answers to these questions necessarily affect the incentives merchant generators will have to make investments. Realistically, especially at this stage of the development of a competitive wholesale electricity markets, SO discretion, as it affects the number and value of transmission rights and uncertain rules for implementing feasibility standards and defining the number of rights introduce uncertainties and potential opportunism problems that are not present for typical property rights.

The impact of a generous ($K\left(q^+\right)$, say) or conservative ($K\left(q^-\right)$, say) distribution of rights on investment incentives depends on the way the resulting shortfall or surplus is financed or redistributed. Suppose, first, that one appeals to the taxpayer. Even if taxation were lump-sum, there would still be distortions in investment behavior; generous distributions over-incentivize merchants, while conservative ones under-incentivize them. By contrast, taxes on users of the transmission network[21] make biased distributions of rights neutral in this radial network, provided that the dispatch is efficient: An efficient dispatch implies that in each state of nature, the cum-tax (or subsidy) price at a given node exhausts the link. And so end-users are unaffected by a generous or conservative distribution of rights on the line. Because there is no source or sink of money outside of the industry, rights owners receive the same overall dividend (for example, through a smaller per-unit dividend in the case of a generous distribution).

Rather than worrying about how to finance/spend the budget deficit/surplus, it seems more natural to simply divide the merchandizing surplus proportionately among the rights' owners. The next box analyzes the optimal *relative* allocation rule (in the same way percentage ownership, but not the total number of shares, matters to determine one's proceeds from the distribution of a firm's dividend, the exact number of rights does not matter as long as the merchandizing surplus is distributed proportionately among rights' owners). It shows that the optimal allocation rule derives from standard asset pricing (CAPM) principles in finance. An addition to an existing link is particularly valuable if its actual capacity remains high when the primary link is very congested. Its construction then creates a diversification benefit.

For example, suppose that the primary North-South AC line exhibits reduced capacity (or breakdowns) during very cold weather. Then an addition along the same path has less social value if it is an aerial AC line than if it is an underground DC line not subject to the same climatic shocks. Allocations of rights in proportion to expected link capacity therefore miss the

[21] We here have in mind proportional taxes on electricity in the South (so the price is $p_S + t_S$, where p_S is the nodal price in the South) and a tax on exports from the North (where generators receive $p_N - t_N$). In the case of $K < K^*$ this effectively reduces the net dividend paid to rights holders so that there is sufficient revenue to compensate them.

point that, for a given capacity, some lines provide better insurance than others. More generally, an allocation of rights in proportion to expected capacities provides insufficient incentives to build lines whose availability covaries with the shadow price less (in absolute value) than that of the existing lines.

Non-contingent financial rights under state-contingent capacities.

Consider the North-South network described at the beginning of the section (see figure 1). The initial expected capacity of the link is K (the actual capacity will in a moment be assumed to be state-contingent). A merchant investor contemplates adding a small amount dK of expected capacity to the line.

Actual dispatching depends on the realization of the state of the world $w.$ The state of the world encompasses the uncertainty about net demand in the South, $D_S(p_S, w),$ that about net supply in the North, $S_N(p_N, w),$ and the actual capacity of the lines :

$$\left[1+q(w)\right]K \qquad \text{for the existing link(s)}$$

and

$$\left[1+m(w)\right]dK \qquad \text{for the new facility,}$$

where we can normalize the noises to have zero means

$$E\left[q(w)\right]=E\left[m(w)\right]=0.$$

Let us assume that the SO dispatches optimally given the state of nature:

$$D_S\left(p_S(w), w\right)=S_N\left(p_N(w), w\right)=\left[1+q(w)\right]K+\left[1+m(w)\right]dK.$$

Let $h(w)\equiv p_S(w)-p_N(w)$ denote the (state-contingent) shadow price of the link.

Suppose further that K^* rights are distributed among all rights owners, including the merchant investor, and that the distribution is proportional to average capacities ; and so the merchant investor receives

$$\frac{dK}{K+dK}K^* \text{ rights.}$$

The merchandizing surplus is distributed to rights owners. Needless to say, distributions of rights that would not reflect expected capacities could by themselves introduce a bias in merchant investors' incentives. For example, suppose that the incumbents in the past received a very generous rating of the existing lines form North to South, while rating standards are strengthened for new comers. The latter then receive a disproportionately small share of total rights, which penalizes them when the merchandizing surplus is distributed among rights' owners, and thereby gives little incentive to sink new investment. To avoid such obvious biases, we assume an allocation of rights proportional to average capacity. Even so, merchant incentives may be inappropriate, as we will see shortly.

The congestion dividend, $d(w),$ paid to the owner of a right is therefore :

$$K^{*}d\left(w\right)=\left[\left[1+q\left(w\right)\right]K+\left[1+m\left(w\right)\right]dK\right]h\left(w\right).$$

The merchant investor’s expected revenue, *R*, is therefore

$$R=E_{w}\left[\left(\frac{dK}{K+dK}K^{*}\right)d\left(w\right)\right]$$

$$=\left(dK\right)E_{w}\left[\frac{\left[1+q\left(w\right)\right]K+\left(1+m\left(w\right)\right)dK}{K+dK}h\left(w\right)\right]$$

$$\simeq\left(dK\right)E_{w}\left[\left[1+q\left(w\right)\right]h\left(w\right)\right]$$

for dK small.

By contrast, the increase in social welfare is

$$dW\simeq\left(dK\right)E_{w}\left[\left[1+m\left(w\right)\right]h\left(w\right)\right],$$

Since in state w, the merchant investor’s expansion delivers $\left[1+m\left(w\right)\right]dK$ units of transmission which each have value $h\left(w\right)$.

Hence,

$$R\leq dW\quad\Leftrightarrow\quad\operatorname{cov}\left(m,h\right)\geq\operatorname{cov}\left(q,h\right).$$

Let us draw the implications of this simple characterization in specific environments :

Example 1 (diversification effect). Suppose that all uncertainty results from line availability. Exist ing line(s) may exhibit reduced capacity due to harsh weather (freezing) conditions. The merchant investor’s line, by contrast is not (or at least less) subject to these harsh weather conditions (or is better protected against them). For example, the new line could be underground, or cross a climatically distinct area. Then $q\left(w\right)$ and $h\left(w\right)$ are (in a first approximation) perfectly negatively correlated, while h is not perfectly correlated with q :

$$m=kq+e$$

with

$$k<1\quad\text{and}\quad E\left(e\left|q\right.\right)=0.$$

Hence:

$$\operatorname{cov}\left(m,h\right)\geq\operatorname{cov}\left(q,h\right),$$

implying

$$R<dW.$$

Non-contingent rights create an under-incentive to invest. Intuitively, the new line supplies a disproportionately high share of the transmission capacity in those states of nature in which transmission capacity is scarce and therefore very valuable. This contribution however is not reflected in the distribution of dividends which is based on fixed (non state-contingent) shares.

It is only when the availabilities of the lines (old and new) are perfectly correlated that the private and social incentives coincide.

The analysis can be generalized to encompass uncertainty about energy market participants's demand and supply curves. Suppose that

$$S_N = a_N + b_N p_N + e_N$$
$$D_S = a_S - b_S p_S + e_S.$$

So $w = (q, m, e_N, e_S)$. Under efficient dispatching

$$h(w) \simeq \left[\frac{a_S}{b_S} + \frac{a_N}{b_N}\right] + \left[\frac{e_S}{b_S} + \frac{e_N}{b_N}\right] - \left[\frac{1}{b_S} + \frac{1}{b_N}\right] K(1+q).$$

This implies that the analysis above generalizes when line availabilities are independent of demand and supply shocks $(\operatorname{cov}(e_i , q) = \operatorname{cov}(e_i , m) = 0$ for i=N,S).

On the other hand, line availability may be related to demand and supply shocks. For example, it may be that a line (old or new) is subject to the same climatic shock as the demand node.Cold weather may simultaneously increase demand and limit the capacity of the line bringing electricity from a cheaper node (precisely when the line is most needed). Such a line obviously has a lower social worth than one whose availability is less negatively correlated with increases in demand at the expensive node.

Example 2 (uncertainty about energy market players only) . Suppose that there is no uncertainty about the actual capacities of the lines :

$$q(w) = h(w) = 0 \text{ for all } w.$$

Hence all uncertainty comes from generation and consumption. In this case, the private and social incentives coincide :

$$R = dW.$$

f) *Objective transmission capacity (2)*

The difficulty in putting a number in front of a line's "capacity" (in the two-node case) raises another issue.[22] As we already noted, rewarding merchant investment through congestion rents requires separating ownership and dispatch in order to obtain an unbiased measure of this rent.

But this separation of ownership and dispatch raises a moral-hazard-in-teams problem. The electric system's state-contingent output (to simplify, the intensity of power in the absence of outage and the probability and duration of an outage) depends on both the care and the forecasts of the owner (the quality of the line, its maintenance, and the adequacy to consumers' needs) and the quality of the management of the grid by the system operator, as the latter must use her acumen to get lots of power through without creating a high risk of outage.

[22] We discuss these issues in the three-node case with loop flow further below.

In other words, the transmission owner's measure of performance is conditioned by the system operator's behavior and therefore incentive scheme. This raises two points: First, one cannot consider incentives given to merchant investors without also specifying those of the system operator. Second, moral hazard in teams reduces accountability. An outage can be claimed to result from poor line maintenance or from imprudent dispatching. Conversely, high power prices may be due to a proper dispatching motivated by low line quality or to an undue conservatism of the system operator.

There is also a potential moral-hazard-in-teams problem among line owners. Recall that merchant investment incentives are better aligned with the public interest when merchants don't have inframarginal units whose congestion rent is to be preserved. The total North-South capacity may then belong to different owners.[23] The same value of a given actual capacity K selected by the independent system operator may correspond to different quality configurations of the different components. The question is then one of allocation of total capacity and congestion rents among the different owners.

Moral hazard in teams : transmission owners and system operator

Consider the North-South network. Let K denote the nominal capacity of the line. In a first step, we assume that this capacity is known to the system operator (for example, the line's maintenance is perfectly observed by the SO). The system operator choose to allow an amount $\hat{K}$ to flow through the link. We assume that with probability $x\left(\hat{K}-K\right)$ the link breaks down and no power flows through it. With probability $1-x\left(\hat{K}-K\right)$, $\hat{K}$ flows through. The function x is increasing.

Let $\mathscr{L}\left(\hat{\mathrm{K}}\right)$ denote the out-of-merit dispatch cost when the realized capacity of the line is $\hat{K}$. We assume that there is no market power at either mode and so $\mathscr{L}$ represents the social loss attached to the inability to import power without constraint from the North. Note that

[23] This is not a necessity, since the incumbent owner, under the threat of new investment by merchant investors, may systematically preempt these investments, and thus there may be a single owner even though the entrants otherwise have more incentive to invest than the incumbent owner.

$$\mathscr{L}' = -h.$$

The socially optimal dispatch solves, for a given K,

$$\min_{\hat{K}} \{ x\left(\hat{K} - K\right) \mathscr{L}(0) + \left[1 - x\left(\hat{K} - K\right)\right] \mathscr{L}\left(\hat{K}\right) \}.$$

And so $\hat{K} = \hat{K}^*$ is given by

$$x'\left(\hat{K}^* - K\right)[\mathscr{L}(0) - \mathrm{L}(\hat{K}^*)] + \left[1 - x\left(\hat{K}^* - K\right)\right] \mathscr{L}'(\hat{K}^*) = 0.$$

The marginal social gain from capacity expansion is then (using the envelope theorem) :

$$\frac{dW}{dK} = x'[\mathscr{L}(0) - \mathscr{L}(K)\big] = (1-x)h.$$

And so, if the marginal investment is rewarded by the congestion cost in the absence of outages, merchant investors face the proper signal for investment.

• *Dispatcher with conservative incentives.*

Turn now to the system operator’s incentives. Suppose that the SO is penalized more for outages than she is rewarded for increases in the amount of power flowing through the network ; that is, she solves :

$$\min_{\hat{K}} \{ x\left(\hat{K} - K\right)q \mathscr{L}(0) + \left[1 - x\left(\hat{K} - K\right)\right] \mathscr{L}(\hat{K}) \},$$

where $q > 1.$ This yields first-order condition :

$$x'[q \mathscr{L}(0) - \mathscr{L}(\hat{K})] + [1 - x] \mathscr{L}'(\hat{K}) = 0.$$

The marginal social gain from capacity expansion is

$$\frac{dW}{dK} = x'[\mathscr{L}(0) - \mathscr{L}(\hat{K})] + [x'(1-q)[\mathscr{L}(0) - \mathscr{L}(\hat{K})]]\frac{d\hat{K}}{dK}.$$

And so

$$\begin{aligned}\frac{dW}{dK} - (1-x)h &= [x'[\mathscr{L}(0) - \mathscr{L}(\hat{K})] + (1-x) \mathscr{L}'(\hat{K})\left[1 - \frac{d\hat{K}}{dK}\right] \\ &= x'(1-q) \mathscr{L}(0)\left[1 - \frac{d\hat{K}}{dK}\right],\end{aligned}$$

using the SO’s first-order condition.

Next, rewrite the SO's optimization program as the choice of a risk factor $\Delta \equiv \hat{K} - K$:

$$\min_{\hat{K}} \{ x(\Delta)q \ \mathcal{L}\ (0) + \left[1 - x(\Delta)\right] \mathcal{L}'(K + \Delta) \}.$$

The cross-partial derivative of the minimand with respect to K and Δ is positive as $\mathcal{L}'' > 0$, and so, by a revealed preference argument, Δ is decreasing in K, or :

$$\frac{d\hat{K}}{dK} < 1.$$

In words, the system operator takes less risk as *K* increases, because the marginal gain from increased throughflow decreases. We therefore conclude that

$$\frac{dW}{dK} < (1 - x)h,$$

and so congestion rent payments over-incentivize merchant investors. In a sense, the SO's conservative behavior implies that insufficient use will be made of the added capacity and so the shadow price of the link overstates the value of additional capacity.

This result shows that one cannot analyze merchant investment (or, for that matter, the incentives of a Tranco company not responsible for dispatching) without considering the system operator's incentives.

Moral hazard in teams : general considerations

At an abstract level, one can view transmission owners and the SO as a team (in the sense of Holmström 1982) jointly delivering an output –state-contingent power– to the final consumers. A general principle is that proper incentives require that each member of the team be made residual claimant for the team performance. So for example each member of a *n*-member team should receive $1 when the team's profit increases by $1 (third parties must act as « budget breakers » to bring the missing $(*n*-1)).

Here, the performance of the team is not a profit, but rather (minus) the social loss

$$x\,\mathcal{L}(0) + \left[1 - x\right] \mathcal{L}(\hat{K}),$$

and the members of the team are the SO and the transmission owners. Making each residual claimant is however very costly for two reasons :

- Adverse selection : fortuitous improvements in performance give rise to *n* rents.

- Collusion : relatedly, the members of the team have an incentive to collude. Suppose for example that a merchant investor has a marginal project that costs as much as the marginal reduction of redispatching cost it brings about. While this merchant investor is indifferent as to whether to implement the project, the other participants (SO, other transmission owners) each costlessly receive the value of this reduction in redispatching cost if he implements it. They therefore have incentives to bribe him into investing. More generally, collusion will induce investments whose cost vastly exceed their social benefit.

To avoid or alleviate these problems, one can make each member accountable for only a fraction of the social benefit. But this policy creates moral hazard. For example, the SO has reduced incentives to dispacth properly (for example, x increases for a given $\hat{K}$) and transmission owners have reduced incentives for maintenance.

g) *Defining and allocating rights with loop flows*

Loop flow introduces additional practical complications. First, how does one define the "capacity" created by new investment and the associated financial rights that go along with the new capacity and a network with three or more nodes and associated loop flows? Second, full or partial outages of one link may affect the effective capacity and nodal prices on other links and at other nodes in less straightforward ways than in the two-node case, especially when there are multiple owners.[24] And as is now well-known, an addition of capacity may have negative social value and even in the absence of system operator discretion, the increase in a link's capacity is unrelated to the system's increased capacity. Finally, and more crucially, small investments may no longer be "marginal".

With a two-node network or a radial network with multiple generation nodes but without loop flow, transmission rights (whether physical or financial) are naturally conceptualised as “link-based” rights reflecting the capacity of each link. When there are more than two nodes and loop flow, then there are at least two ways of introducing financial rights (Joskow and Tirole 2000, pp. 478-479). One approach is to use “link-based” rights (Oren et. al. 1995) which are rights associated with each transmission line on the network and, in the case of financial rights, paying a dividend equal to the shadow price of the congestion on each line. The other approach proposed by Hogan (1992) is to specify point-to-point financial rights from each injection node

[24] For example, there are simultaneous import limitations into California that depend on the availability of links from the Southwest to Southern California, the Northwest to California, and the operating generating capacity inside California. These limits are presently managed administratively with “nomograms” that define the curtailments that are triggered when the constraints are binding.

to each receipt node on the network, with the rights paying a dividend (which could be negative) equal to the difference in nodal prices at the two nodes due to congestion. We will focus on point- to- point financial rights here since they are being used in several areas of the U.S. and appear to be the favored approach in the FERC's SMD rules. Moreover, in theory the values of point-to-point rights internalise network externalities associated with loop flow since they reflect the shadow prices on all lines affected by an injection at one node and an equivalent withdrawal at another node. The shadow price on a particular transmission link, however, does not reflect the social value of the link to the network overall.

For the general case of a multi-node network with loop flow, Hogan (1992) envisions that point-to-point transmission rights will be defined and allocated through a process in which a set of all feasible (i.e., consistent with the transmission network) *physical* combinations of bilateral contracts between injection and receipt points is first calculated. The process of defining the feasible set must be conducted by the SO by performing a large set of simulations of the use of the network under various supply and demand conditions and contingencies (e.g. line outages) using load flow models. The process envisioned for defining the feasible set appears to be purely physical in the sense that the SO does not rely on prices or other valuation procedures to define the set of feasible rights. A second process (e.g. grandfathered allocations to incumbents, auctions, bilateral trading) is then used to define the specific combination of rights/capacities from within (or on the frontier of) the feasible set that will be allocated initially to generators, marketers and/or load serving entities. Once a specific combination of feasible transmission rights is defined and allocated they become the property of the holders. These rights may then be traded in secondary markets.

Investments in new transmission capacity are translated into *incremental* transmission rights through the effects these investments have on *shifting* the initial frontier of the set of all feasible bilateral transactions and the associated configurations of point-to-point capacities/rights. The literature generally assumes (a) that the initial feasible set and shifts in its frontier are well defined in the sense that there is no uncertainty about the relevant parameters of the feasible set, (b) that the feasible set does not itself vary with exogenous random variables, and (c) that the shifts in the frontier of the feasible set do not make any rights/capacity combinations that were previously in the feasible set infeasible, post investment, or else that efficient trading arrangements are in place that ensure reallocations to ensure feasibility.

There are both practical and theoretical issues that may undermine these assumptions. As we have already discussed, the feasible set of bilateral schedules that can be accommodated without causing congestion and the associated transmission capacities and accompanying rights depend on exogenous environmental parameters. While this fact literally contradicts assumption b) above, the existing theory can straightforwardly be extended in the usual manner by allocating *state-contingent rights*, as long as the contingencies can be described (temperature, output of specific generators that affect contingency limits,conditions in interconnected control areas, etc.). The drawback of this extension is that the large number of potential contingencies that are relevant for defining and implementing the feasibility requirement call for a large number of state-contingent rights, with the concomitant problems that these create: large transaction costs, thinness and market power in the secondary markets for these rights.

Assumption a) can also be questioned. In particular, system operators have substantial discretion on defining and implementing security constraints, affecting the actual power flows on the network in real time, and random line outages. Moreover, for complex networks the physical feasibility evaluation necessary to define the numerous potential configurations of transmission rights that are simultaneously feasible and the incremental configurations of transmission rights created by new investments involves many discretionary assumptions and is likely to be based on

DC-load flow models that are approximations to real networks, are subject to SO discretion and may not be especially good approximations under stressed conditions when losses are significant and contingency constraints binding. These are the conditions when transmission rights are likely to be especially valuable. What is modeled as being feasible and what is feasible in actual operations can differ, especially when reactive power and voltage constraints are important.

It should be clear that the merchant transmission model cannot operate “as if by an invisible hand,” since some *de facto* regulatory authority must have the ability accurately to simulate load flows on the network, apply contingency criteria, define feasible sets and changes in feasible sets associated with transmission investments, and ensure that rights allocations are consistent with feasibility under numerous contingencies.

Let us finally come to assumption c). It is clear that, except in radial networks, the expansion of the network both creates new feasible allocations and makes some initially feasible allocations infeasible. So, in general, the expansion may infringe on existing property rights. This problem has been recognized in the academic literature, though not very clearly in the policy arena, and it has been proposed that the merchant investor building a new line leave existing property rights intact, which in general requires the merchant investor to compensate for the loss of property rights by buying existing ones and turning them back to initial owners who were expropriated. These issues are illustrated in the box below.

Network expansions and infringements on point-to-point financial rights

The most elegant explanation of how the contract network framework can be applied in practice to a simple network with loop flow is provided by Bushnell and Stoft (1997), and we follow their presentation very closely here. Figure 3 depicts the standard simple three-node network. There is generation in the North, generation in the South and demand in the East. The transmission lines connecting these nodes have capacities K_{NE}, K_{SE} and K_{NS} respectively as depicted in Figure 3.

Assuming that the transmission links connecting the three nodes are of equal length (resistance) and ignoring losses, the physical laws of electricity (Kirchoff's) determine the flows through the three transmission links associated with alternative configurations of generation (q_N and q_S) in the North and South and consumption (q_E) in the East. The relevant constraints applicable to the definition of the feasible set of bilateral transactions are:

Figure 4

The feasible combinations of q_N and q_S associated with each constraint are depicted by lines in Figure 4 and the intersection of these sets defines the feasible set of bilateral transactions.[25] The feasible set is depicted as the hatched area in Figure 4 (equivalent to Figure 2 is Bushnell and Stoft 1997). The dispatch of the system and the allocation of point-to-point transmission rights must lie within this feasible set.

Accommodating investment into this framework is tricky because grid expansions can both make

[25] In what follows, the K_{NS} is assumed to be small relative to K_{NE} and K_{SE}. The capacities K_{SE} and K_{SE} do not have to be equal, but the examples in Bushnell and Stoft (1997) assume that they are and we will follow that assumption in the graphical presentation here.

combinations of q_N and q_S and associated point-to-point rights feasible that were not previously feasible and make some combinations of q_N and q_S and associated point-to-point rights that were feasible pre-investment, infeasible post-investment. To see this, it is again useful to follow Bushnell and Stoft (1997) and start with a radial network that does not have a link connecting the North and the South, and, accordingly, no loop flow. A radial network of this type is depicted in Figure 5.

For this network, the feasible set is very simple to define. It satisfies:

$$\begin{aligned} q_N &\leq \mathrm{K}_{NE} \\ q_S &\leq \mathrm{K}_{SE} \\ q_N + q_S &= q_E. \end{aligned}$$

Generation at each node can is limited only by the capacity of the link connecting it to the demand node. The feasible set of bilateral transactions for this radial network is depicted as the hatched area in Figure 6. Figure 6 also depicts a hypothetical optimal dispatch consistent with generation in the North being less costly than generation in the South, but limited transmission capacity from North to East requires some more expensive generation from the South to be dispatched to clear the market.[20] The marginal cost of the expensive generation that clears the market would also determine the market clearing price in the East.

Now let's consider adding to this radial network a third link between North and South to create the three-node network with loop flow depicted in Figure 3. The changes in the feasible set resulting from this investment are displayed in Figure 7.

(This figure is equivalent to Figure 4 in Bushnell and Stoft 1997.) Some allocations that were previously feasible are now infeasible and some allocations that were not previously feasible are now feasible. In particular, the initial optimal dispatch is no longer feasible. In order to go forward with the new line, the investor would (effectively) have to buy back sufficient rights from those who hold them initially to restore feasibility (or as in Bushnell and Stoft require the investor to take rights that have negative values and require payments rather than receiving dividends to restore feasibility). An efficient economic transmission rights reallocation process must complement any physical analysis of the effect of a transmission investment on the feasibility of the existing allocation of rights. If the SO were to take the allocation of existing rights as fixed when performing a feasibility test, the set of investments that satisfy the constraint that no existing right will be made infeasible will lead to a set of allowable investments that is much smaller than the set of investments that increase social welfare. With such an efficient reallocation mechanism in place, Bushnell and Stoft show that that it will be most profitable for the investor to acquire rights that lead to an allocation equal to the most efficient dispatch given the constraints associated with the new investment and associated network topology. Using their numerical assumptions, the new efficient dispatch and allocation of point-to-point financial rights would be the point depicted in Figure 7 that involves less (cheap) generation in the North and more (expensive) generation in the South than was the case without the new link. The new link is therefore inefficient and should not be built and, indeed, Bushnell and Stoft show that the obligation of the investor to restore feasibility will make this investment unprofitable for a merchant investor under these assumptions.

This naturally raises the question of why transmission links such as the one between North and South that cause loop flow are so common. One reason might be that the post-investment optimal dispatch lies in the new feasible region, allowing increased production from the cheap generator in the North. This could be the case, for example, when demand is high and the optimal dispatch for the radial network is further to the right in on the K_{NE} constraint in Figure 7, involving more generation from (expensive) South. The new link would then have the effect of increasing the feasible (cheap) supplies from North and reducing the (expensive) supplies from South to balance supply and (higher) demand, by effectively increasing the capacity from North to East via South. But, this situation requires that it is less costly to invest in transmission capacity to increase supplies from the North over an indirect path (North to South to East) than simply to increase the capacity of the direct link from North to node East.

If the new link does not move the efficient dispatch into the new areas of the feasible set, the third link between generation nodes in the standard three-node network reduces social welfare because it is a binding constraint on low-cost generation schedules which would otherwise be accommodated without congestion on the direct links between each generation node and the demand node. This link only makes sense when we recognize that one of the other links may fail and the third link provides an alternative path for delivering supplies to satisfy demand (Joskow and Tirole 2000, p. 477). So, this link will have negative value under some contingencies and positive value under others. Adding the link will reduce the feasible capacity from at least one node to another under some contingencies and increase it under others.[21]

We can see this by extending the Bushnell and Stoft's (1997) examples to take account of line outages. Let's go back to the radial network depicted in Figure 5. Assume now that that the capacity of the link from North to East is reduced as a consequence of equipment failures or other contingencies; assume that the capacity is cut by 1/3. The new feasible set is depicted in Figure 8.

The new optimal dispatch involves less cheap generation and more expensive generation and increases total generation costs. Now we consider the effects of adding a link between North and South. The new feasible set (under the condition that capacity on the link between North and East has been cut by a third) is depicted in Figure 9.

As before, adding a line that causes loop flow makes some allocations that were previously feasible now infeasible and others that were previously infeasible are now feasible. As portrayed in Figure 9, the link in this case makes it possible to increase generation at (cheap) North and reducing generation and (expensive) South . Accordingly, the third link will reduce generation costs (and perhaps reduce the probability that demand will have to be shed to balance supply and demand) when there are transmission line outages of the type examined here. Whether it is an efficient investment will depend on the benefits of the link during contingencies like these (and others when it is valuable), the costs of the link during conditions when it is not "needed" and leads to an inefficient dispatch, and the cost of the investment. However, it is not clear whether or how the number and allocation of *non-contingent* transmission rights can be defined to capture the varying valuations of a transmission investment under the many contingencies that characterize real electric power networks and provide the right incentives to support efficient investments. Moreover, for any mechanism like this to work well a liquid competitive secondary market for rights would have to exist to make it possible for investors to easily buy and sell rights at their competitive market values to restore feasibility and to allow welfare enhancing investments to go forward.

The success of any property-based system in attracting efficient levels of investment depends on the ability to define and enforce clear and consistent property rights. This appears to be an especially challenging problem on an electric power network with loop flow where the feasible set of property rights and their efficient allocation (i.e. not just their value) are contingent on changing supply and demand conditions, the application of contingency constraints by the system operator, [22] and their interaction with new investments.

g) *Forward markets and commitment*

As we have seen, merchant investment is most appropriate for new investments. Constructing a new line, however, involves both a long lead time and substantial uncertainty as to the availability of a crucial input, namely the various authorizations needed to build the line, and as to the nodal prices of electricity in the distant future. This gives rise to three concerns:

- *Availabiltiy of financing*.

Merchant investment is a high-powered-incentives activity. Merchants thus bear a substantial long-term risk. To obtain financing, they probably will want to unload a good part of this risk.[26] One technique for doing so consists in entering financial arrangements with generators and load-serving entities. The latter then still face energy price risk as well as (if this transmission project cannot be brought to completion) counterparty risk. In principle, some insurance should also be supplied by non-stakeholders. For reasons that have received insufficient attention in economic theory, such long-term forward markets are usually poorly developed, though. This fact can make it hard for merchant investors to raise financing.

- *Credibility vis-à-vis projects with shorter lead times*.

[26] For theoretical foundations for the desirability of this unloading, see Holmström-Tirole (2000).

Transmission projects compete with generation ones. Suppose for instance that a merchant investor plans to invest in a new North-South line, whose construction will take about 10 years, and faces a rival generation project in the South, that takes only 2 years to build. There is room for only one of these two alternative ways of reducing the price wedge between North and South. The merchant investor is at a strategic disadvantage even if his project is socially more valuable. If most of the costs involved in building a new line are sunk after the first two years, then the merchant investor is likely to cancel his project if the new generation plant in the South is built. Knowing this, the generator may well try to use his short-term investment period to preempt the transmission project, and this even if the merchant investor has announced his intention and has started work on this project.[27]

- *Regulatory uncertainty and opportunism*.

Government and regulators have substantial discretion over the profitability of energy projects. In the case of the construction of a new line, they will first affect the probability that the company receives the authorizations needed to build it. And, once it is built, the choice of rating paradigm (which determines the number of rights allocated to the merchant), the imposition of energy price caps, the definition of incentives for the System Operator (see (f) above), the build-up of parallel lines under different incentives (e.g., by a Transco regulated under cost-of-service and aiming at reducing nodal price differences or market power in the South) all impact the merchant investor's long-term return.

While this commitment problem exists for all investments, it is partially mitigated on the short end by institutional factors (short-term stability of the regulatory environment, 5-year regulatory commitments, …) and by the current regulators' reputation concerns. But long-term commitments are less desirable and administrations change. This is why long term investments

[27] Such timing issues are of course not specific to transmission investments. But the latter are particular vulnerable

whose payoffs are heavily dependent on government policies are often performed either by a State-owned enterprise or by a utility under some cost-of-service scheme, but not by a private company under a high-powered incentive scheme.[28]

4. Regulated transmission company

It is quite clear that a pure merchant transmission model could lead to potentially serious inefficiencies as a consequence of a number of significant market imperfections that are likely to be created by the economic attributes inherent in transmission investment and associated network effects. Of course, most real markets are characterized by at least some imperfections. Why should such market imperfections be of particular concern with regard to transmission investment? We believe that there are at least two sets of reasons why these potential market imperfections should be of concern. First, there is an unusual combination of imperfections associated with economies of scale, scarce transmission corridors, network deepening investments, missing markets, non-price rationing, definitions of credible property rights, and generator market power together suggest that transmission investment is less likely to be conducive to efficient governance by market mechanisms than is the case for most unregulated markets; the imperfections are attributes that are typically used as justifications for regulated "natural monopoly." Second, the experience with liberalized power markets to date suggests that performance of these markets deteriorates significantly when there is even a small degree of under-investment in transmission and generation investment. Because electricity is non-storable, is characterized by very inelastic short run demand, and the associated markets operate subject to physical and reliability constraints that are difficult to fully capture in prices determined in

to preemption strategies due to their long lead time.

[28] Unless legal protection against expropriation may be supplied by the court system, which requires that expropriation take blatant, rather than subtle forms.

market clearing processes, a small amount of under-investment can have a large impact on congestion and market prices, significantly increase market power problems, and increase the need to use non-price administrative mechanisms to balance supply and demand in real time when markets cannot clear fast enough to do so or market power mitigation mechanisms kick in. On the other hand, when there is little congestion abundant capacity competitive power markets appear to work quite well. Accordingly, the social costs of equivalent amounts under-investment and over-investment in transmission are likely to be asymmetrical; the social costs of too little are higher than the social costs of too much.

There are two alternatives to a pure merchant investment framework for governing transmission investment. The first is to rely on regulated monopoly Transcos (as in England and Wales, Spain and a number of other countries)[29] to be responsible for transmission investment.[30] The second is to have a system in which both merchant investments and regulated investments are accommodated as in Australia, somehow combining the merchant and regulatory governance framework (hopefully) in a complementary fashion.[31] Of course, designing a perfect regulatory (or combined) system is not possible either. In the end, to choose between alternative institutional frameworks one must compare the likely costs of the imperfections of each to determine what is the best that we can do in an imperfect world.

[29] In England and Wales the National Grid Company (NGC) is and has been both the TO and the SO since privatization, but this reflects a decision by the regulator based on ongoing assessments of the costs and benefits of separating TO and SO function, rather than a legal requirement that these functions be integrated.

.

[30] One can't really avoid issues associated with regulating Transcos since all liberalized systems start with legacy transmission networks that have been subject to regulation historically and will continue to be subject to regulation for the foreseeable future.

[31] Regulators and politicians are often attracted to this alternative because it seems to make everyone happy; or you can have the best of both. We caution about jumping to such a conclusion. Mixing regulation and competition efficiently is very challenging and the experience with this model in Australia is not encouraging unless maximizing litigation is viewed as being socially valuable. A model that combines regulated and merchant transmission is likely to work best if some way can be found to distinguish between investment opportunities that are most conducive to a merchant model (e.g. DC lines between separate market areas with large differences in competitive power costs) and those that are most conducive to a Transco model (e.g. network deepening investments).

There has been surprisingly little serious research done on the design of economic regulatory mechanisms for Transcos.[32] This is surprising if for no other reason that we have been regulating the operation of, investment in, and prices for transmission networks (as segments of regulated vertically integrated firms) for almost a century. Regulatory issues associated with transmission probably attracted so little attention because transmission costs (capital and direct operating) are a small fraction of the total bundled price of electricity (typically 3% to 8%), transmission investments were typically made by vertically integrated firms in conjunction with investments in new generating plants which were intended to run without being constrained by transmission congestion, there was the perception that at least within control areas congestion was not a serious problem,[33] and competitive power market performance problems created by transmission congestion were not an issue for regulated vertically integrated monopolies relying on cost-based dispatch rules and reliability protocols and rewarded with cost-based prices.

Nevertheless, we can draw on the extensive literature on regulatory mechanisms, and their application and associated performance in a variety of contexts (including electricity) to articulate what a good regulatory mechanism would look like and what its strengths and weaknesses are likely to be. We know from this literature that there are several issues that a good regulatory mechanism for Transcos will have to address. First, the regulatory mechanism must satisfy the constraint that the Transco can earn revenues that are at least high enough to cover its capital and operating costs. Second, the regulatory mechanism should provide incentives to the regulated firm to make efficient operating and investment decisions. Third, the

[32] Celebi (undated), Nasser (1997), Léautier (2000), and Vogelsang (2001).

[33] On the other hand, the focus of vertically integrated utilities on their own control area facilities combined with biases toward owning generating plants rather than buying power from third-party suppliers in neighboring systems probably led to under-investment in transmission facilities connecting control areas and to some of the problems we now see with trans-border wholesale power trade in Europe and interregional trade in the U.S.

regulatory will have imperfect information about the cost opportunities facing the firm and the value (demand) for the services it provides. The regulator can adopt various practices to improve the quality of its information, but it will never be perfect and, in general, the regulated firm will have superior information to that available to the regulator. Finally, this asymmetry of information available to the regulator and the regulated firm creates a tradeoff between providing good incentives to the regulated firm and providing it with enough revenue to recover its costs (rent extraction). Incentive regulation mechanisms are designed to make the best use of the information that the regulator does have to define an incentive regulatory mechanism that yields the most attractive tradeoff between incentives and rent extraction.[34]

In the merchant investment model, investors are compensated for their costs by receiving congestion rents (or by selling rights to others to receive those rents). For a regulated Transco, however, it is clear that one thing that we do not want to do is to adopt a regulatory mechanism through which the Transco's compensation varies directly with the congestion rents or merchandizing surplus (hK) associated with congestion on its network.[35] This will simply give the transmission owner the incentive to reduce transmission capacity to increase congestion prices. Accordingly, we rule out regulatory mechanisms that have the property that the Transco's compensation increases with congestion rents.

More appealing schemes attempt to confront the regulated Transco with some measure of the social gain or loss associated with its activity. For example, we want to reward the Transco for reducing congestion (efficiently) and penalize it for (inefficiently) increasing congestion costs. Nasser (1997) and Léautier (2000) propose a regulatory mechanism that is a direct application of Laffont and Tirole (1993). It assumes first, that the Transco is presented with a

[34] See Laffont-Tirole (1993).

[35] Unless it is just a credit toward covering the full cost, as in Chile.

menu of revenue sharing contracts that induces the firm to select the contract that provides it with the most (second-best) efficient combination of incentives to control the costs of maintaining and building transmission lines while allowing it to earn enough revenue to at least recover its costs, consistent with the regulator's information. This is complemented by a separate mechanism that provides incentives to the firm to invest in the optimal amount of transmission investment. This mechanism effectively involves placing the Transco at risk for all costs of congestion and any investment costs incurred to relieve it. The mechanism gives the Transco the incentive to minimize the sum of the expected costs of congestion, losses, operational changes and investment. This simple result depends on the applicability of Laffont and Tirole's (1993) dichotomy between the menu of contracts required to solve the rent/extraction efficient tradeoff and the provision of congestion management (including investment).

In a related contribution, Vogelsang (2001) proposes a Transco regulatory mechanism that relies on a two-part price cap. It is easiest to understand the proposal in the two-node case. The first part of the price cap contains a fixed fee charged to each load serving entity on the network (perhaps varying with the number of customers it serves). The second part of the price cap rewards the Transco based on the amount of power flowing over the transmission link. If the link is congested, the Transco can increase quantities by investing to increase transmission capacity. Total revenues from both components of the mechanism are subject to the price cap (set at a level to allow total cost recovery in expectation) and the Transco can choose the fixed fee and the usage fee subject to the overall cap. The Transco will set prices for usage high when there is congestion and low when there isn't, but must reduce the fixed fee to stay within the revenue cap. It is then profitable for the Transco to make investments to expand transmission if the reduction in congestion costs is greater than the cost of the investment. The properties of price cap mechanisms like this one depend on choosing the correct weights, which may be especially challenging for a complex transmission network which, in the traditional price cap framework, is a multiproduct firm with a lot of products. The author recognizes as well that it would have to be adapted to a regime where a security constrained bid-based dispatch must be used to set nodal prices that would not be subject to the Transco’s discretion.

Price caps for Transcos

There are two different forms of price caps: node-based and link-based. Nasser (1997) and Léautier (2000) consider link-based caps .

• *Link-based caps*

$$\sum w_k h_k \leq \text{ constant}$$

(*k*= links). A tax on electricity (possibly node-contingent as in Nasser and Léautier) is then used to cover the costs of the Transco. (For example, if investment eliminates congestion altogether, the constant is equal to zero). The weights are presumably the flows through the links.

There are a couple of issues with this mechanism One, which probably can be handled, is the treatment of interconnectors (links with other systems). A priori, these links must be included in order to give incentives to invest in them, hoping that there are nodal prices in adjacent systems. In a similar spirit, one must treat non-existing links in a proper manner (they have a fictitious shadow price equal to the difference in nodal prices). And there are lots of non-existing links. What weights should be put on these links?

Second, the regulator has no clue as to the right choice of weights and cap. [This is true for any price cap, but the problem seems particularly severe here]. In the North-South example, choosing the cap amounts to choosing an investment (a random one if the regulator makes mistakes in forecasting supply/demand).

Third, the "prices" h_k are equilibrium determined and not chosen by the Transco. So punishments for violations of the price cap constraint must be designed.

• *Node-based caps*

$$\sum w_k p_k \leq \text{ constant}$$

This type of cap (with the weights related to the net supplies/demands at the nodes) avoids some of the problems faced by link-based caps. Still, such caps seem hard to design. Note that the practice of having the transmission-owner pay for must-runs is in the spirit of node-based caps.

Another issue with price caps is related to sunk investments by users (generators, LSEs, industrial users). Standard Ramsey theory (which offers some foundation for price caps) assumes that 1) the utility sets prices and 2) users react by choosing their demand. Here there is a "stage 0" at which the users already sink some investment (this problem is much more general than just electricity).

More generally, there are at least two ways to give a Transco the correct investment and operating incentives (ignoring the overall cost compensation constraint for now). The first is a *surplus-based scheme*. In figure 10, capacity K_0 avoids out-of-merit redispatch cost equal to the area of GADH. An increase to K_1 raises this amount by the area ABCD. Thus the regulated Transco can be rewarded on the basis of the surplus S_1 created by the investment (or the operational improvement). Alternatively, and equivalently, one may consider the redispatch costs AOD and BOC under capacities K_0 and K_1. This *redispatch cost* measure of course leads to the same assessment S_1 of the performance improvement.

The benefit of these schemes is that the Transco then faces the entire social cost of congestion. Lumpiness is no longer an issue. Neither is loop flow. Neither are the special attributes of network deepening investments, transmission rights specification and allocation, preemption and other problems identified with the merchant investment framework. And, as long as the cost of outages is properly accounted for in the redispatch cost (a strong assumption given current practice), there is no problem in putting together the two tasks of Transco and SO, thus eliminating the moral hazard in teams problem that naturally arises under merchant investment.

However, there is no free lunch here. There are limitations to the Transco approach and we begin to explore them below.

a) *Repressed supply/demand*.

In the absence of uncertainty, the investment by generators would be no larger than what matches existing transmission capacity. To see this, return to figure 10 and consider the extreme case in which the demand and supply curves are perfectly inelastic in the short run and there is no uncertainty about demand, generation and transmission availability. Then capacities are perfectly adapted to demand. Under truthful revelation, the bids into the balancing market would reveal for example a net supply curve GAF in the North, and not the "potential" or "long-term" supply curve GG'... So the redispatch cost is always equal to zero. More generally the redispatch cost is likely to understate the missing surplus that is foregone by not investing.

Probably the solution to this problem is to forget about the redispatch cost and simply reward the Transco on the basis of the increase in social surplus from, say, GADH to GBCH.

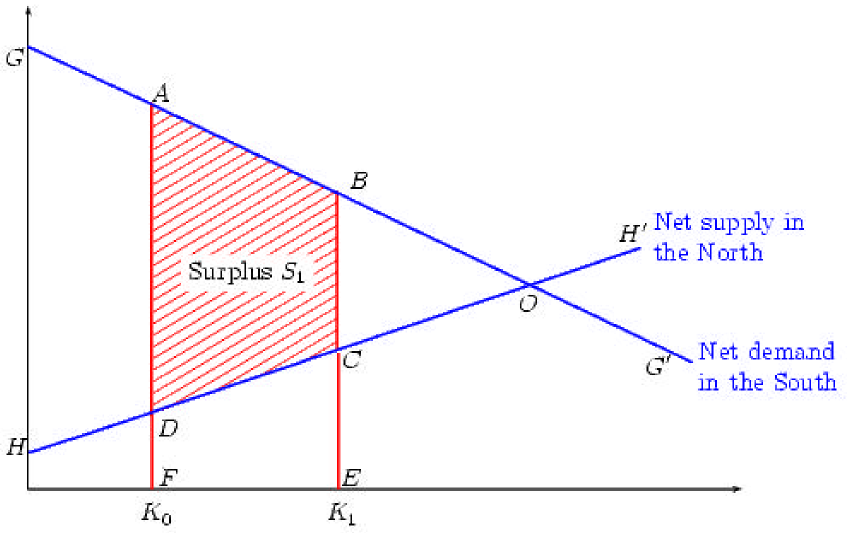

Figure 10

More generally, the measurement of supply and demand curves through the spot market understates the redispatch cost and overstates the surplus created by existing capacity, as shown in figure 11.

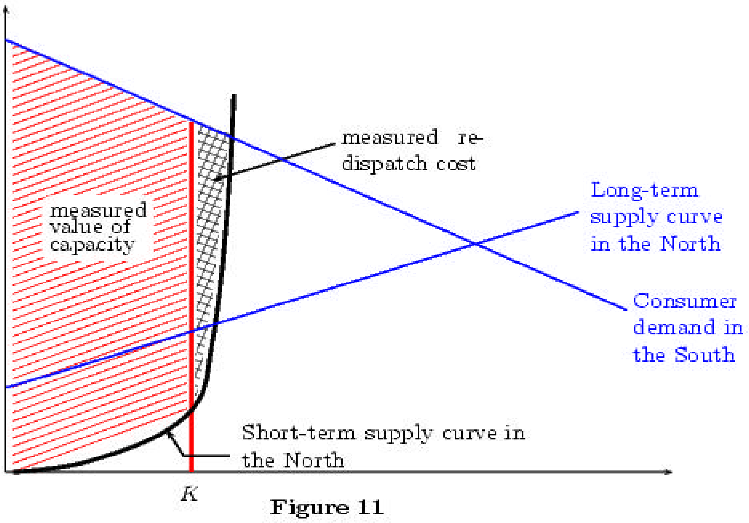

Figure 11

In figure 11, we assume, for simplicity, that all generation is in the North, and all consumption in the South, and that consumers have no investment opportunity (no bypass, no mobility) and so their short- and long-term demand curves coincide. By contrast, the short- and long-term supply curves in the North are very different. Installed capacity, which to a large extent reflects the ability to export through the constrained link, has variable cost way below total cost (indeed, some plants, such as nuclear plants, that must run are likely to bid zero); the short-term supply curve however becomes very steep beyond the line's capacity.

A distinct reason for why surplus measurement is likely to be imperfect is that residential and small business users (and therefore, indirectly load-serving entities) do not bid downward-sloping demand curves. Transaction costs for the moment still prevent systems from having precise estimates of the value to residential users (that is why they rely on rough estimates of "value of lost load" in the case of breakdowns).

b) *Manipulation of bids*.

If the Transco is vertically integrated with generation (which may be desirable for other reasons), the Transco may initially keep old plants in activity (assuming that the incentive scheme is based on the reduction in the uplift).[36]

c) *Thin spot energy markets*.

Bilateral (physical) trades a priori make it hard to have precise information on surplus since willingnesses to pay and ask prices are not observed. "Inc" and "dec" bids, though, restore some of the informational content lost by taking the trade's bid and ask out of the pool.

d) *Rent extraction*.

The issue is here the same as for any regulated company. The uplift may be reduced "by chance" (investment in generation in the South, demographic evolution,...). Similarly, the Transco may have private information. There is no theoretical novelty here; the uncertainty impacts the slope of the incentive scheme.

e) *Coordination with generation and consumption investments*.

Under this scheme, the Transco has an incentive to *understate* its investments in transmission in order to induce users to invest, resulting in an uplift reduction that is costless to the Transco. Conversely, the Transco would like to have information about lumpy investments by the users. Presumably it will get some through new connection requests although these may i) be non-binding and ii) be made at the last minute. Generators have an incentive to overstate the size of their new investments if there is no charge for them .

Transco regulation in England and Wales[37]

When the electric power system in England and Wales (E&W) was restructured and privatised in 1990, an independent Transco (National Grid Company or NGC) was created to own, operate, and invest in the E&W transmission network.[38] NGC is both a regulated TO and a regulated (for-profit) SO. [39] Initially, NGC was subject to a revenue cap regulatory mechanism that allowed its revenues to grow with inflation less a productivity factor (RPI-X). The revenue cap was set as part of a process in which NGC had to submit a five-year plan for future investment and operating costs to the regulator. The plan included forecasts of demand growth, generation capacity additions, security constraints, etc. The proposed plan was subject to public comment and consultation, including advice received from independent consultants retained by the regulator. A final agreed upon five-year plan for capital and operating costs was then agreed to

[36] Even if the Transco is not vertically integrated, the Transco may have an impact on the timing, quality, and capacity of interconnectors with other systems, and of connections of new generators.

[37] This discussion refers to the system in England and Wales during the 1990s, prior to the introduction of NETA in early 2001. See Richard Green (1997).

[38] NGC's exclusive license does not extend to the facilities that connect particular generators to the network itself.

[39] Technically, the TO and SO functions are separable in England and Wales. However, the regulator has historically assigned both functions to NGC in England and Wales. A separate organization ran the pool and the associated settlements system.

with the regulator and an annual revenue cap (RPI-X) mechanism tied to the forecast capital and operating costs applied to NGC. In addition, service quality criteria were specified and financial penalties applied for failing to meet them. The costs (revenues) were then allocated through a variety of charges assessed to generators and distribution companies based, in part, on their locations. After 5 years the process was restarted with a new plan and a new revenue cap (5-year ratchet) put in place.

As an SO, NGC is responsible for managing congestion, providing for operating reserves and frequency regulation, and balancing supply and demand in real time to maintain frequency and voltage requirements. Under the system in place during the 1990s, all generators were required to make bids to supply to the Pool, a separate entity from NGC. The Pool then took the supply bids and various physical parameters of the generators (e.g. ramp rates) and calculated a "least cost" dispatch for each half-hour of the following day. The highest bid accepted to clear the market in each half-hour interval determined the market clearing price for that interval and all suppliers selected to supply were paid this uniform market clearing price.

This dispatch process ignored congestion and losses and the pool price did not vary to reflect those. It was then the responsibility of NGC as the SO to manage congestion and losses and otherwise to maintain the physical reliability and technical attributes of the network. NGC managed congestion by paying generators to increase or decrease their output as necessary to bring physical supply into balance with demand at every node on the network taking the networks physical constraints into account. Basically, a generator whose output was increased was paid an amount equal to the bid that it had submitted to the Pool to supply that level of output. This price would have been higher than the uniform market-clearing price determined in the pool. A generator whose output was reduced to manage congestion was paid the difference between the market clearing price and its bid. In each case, when there were multiple generators available to increase or decrease output to manage congestion, NGC was supposed to pick the lowest cost option. The total costs incurred by NGC to manage congestion in this way constituted its "congestion costs" and were included in an uplift charge that was added to the wholesale price of electricity.

During the four years following restructuring and privatisation congestion management costs tripled. The regulator ultimately decided to impose an incentive regulatory mechanism on NGC as SO to give it better incentives to reduce congestion (and other uplift costs which we will not discuss further here.). The mechanism established an annual budget for NGC's congestion management costs and its actual congestion management costs were then compared to this pre-determined budget. NGC was responsible for a share of costs that exceeded the budget and was able to keep a share of the costs if they were less than the budget. Both the rewards and the penalties were capped. The congestion cost budget was updated annually. This SO incentive scheme was designed to be a complement to rather than a substitute for the planning process and revenue cap mechanism applicable to NGC's TO functions. The SO scheme was designed to improve incentives for day-to-day operations and investments with short-paybacks. The TO scheme (with a 5-year ratchet) provided the financial support for longer term investments. However, since to SO and TO functions were combined to NGC, there were likely to have been some interactions between the two.

The Transco model and the regulatory mechanism applicable to NGC are generally viewed as having been successful. Total transmission charges, congestion and other uplift charges have all declined significantly since the SO incentive scheme was introduced in 1994.

REFERENCES

Baldick, R. and E. Kahn (1992) “Transmission Planning in the Era of Integrated Resource Planning: A Survey of Recent Cases,” Lawrence Berkeley Laboratory, LBL–32231.

Bohn, R.E., Caramanis, M.C., and F.C. Schweppe (1984) “Optimal Pricing in Electrical Networks over Space and Time,” *Rand Journal of Economics*, 15 (3): 360–376.

Bushnell, J. (1999) “Transmission Rights and Market Power,” *Electricity Journal*, 12(8): 77–85.

Bushnell, J. and S. Stoft (1996) “Electric Grid Investment Under a Contract Network Regime,” *Journal of Regulatory Economics*, 10: 61–79.

Bushnell, J. and S. Stoft (1997) “Improving Private Incentives for Electric Grid Investment,” *Resource and Energy Economics*, 19: 85–108.

Celebi, M. (undated) “An Analysis of Incentives to Provide Line Capacity and Reliability in Deregulated Power Networks,” mimeo, Brattle Group.

Chao, H. P. and S. Peck (1996) “A Market Mechanism for Electric Power Transmission,” *Journal of Regulatory Economics*, 10 (1): 25–59.

Glachant, J-M, and V. Pignon (2002) “Nordic Elecrtricity Congestion’s Arrangement as a Model for Europe: Physical Constraints and Operators' Opportunism,” mimeo.

Green, R. (1992) “Contracts and the Pool:The British Electricity Market,” mimeo, Department of Economics, University of Cambridge.

----- (1997) “Transmission Pricing in England and Wales,” *Utilities Policy*, 6(3): 185-93.

Hansmann, H. (1996) *The Ownership of Enterprise*, Belknap Harvard.

Hogan, W. (1992) “Contract Networks for Electric Power Transmission,” *Journal of Regulatory Economics*, 4: 211–242.

----- (2002) “Financial Transmission Rights Formulations,” mimeo, Harvard University.

Holmström, B. (1982) “Moral Hazard in Teams,” *Bell Journal of Economics*, 13: 324–340.

Holmström B. and J. Tirole (2000) "Liquidity and Risk Management,” *Journal of Money, Credit and Banking*, 32 (3): 295–319.

Joskow, P. (2002) “Electricity Sector Restructuring and Competition: A Transaction Cost Perspective,” in Eric Brousseau and Jean-Michel Glachant (eds.) *The Economics of Contracts: Theories and Applications*. Cambridge: Cambridge University Press.

Joskow, P., and R. Schmalensee (1983) *Markets for Power*. IT Press.

Joskow, P. and J. Tirole (2000) "Transmission Rights and Market Power on Electric Power Networks," *Rand Journal of Economics*, 31(3): 450–487.

Laffont, J-J. and J. Tirole (1993) *A Theory of Incentives in Procurement and Regulation*, MIT Press.

Léautier, T. (2000) "Regulation of An Electric Power Transmission Company," *Energy Journal*, 24(1): 61-92.

Nasser, T.O. (1997) *Imperfect Markets for Power: Competition and Residual Regulation in the Electricity Industry*, PhD Dissertation, MIT Department of Economics.

Oren, S. (1997) "Economic Inefficiency and Passive Transmission Rights in Congested Electric Systems with Competitive Generation," *The Energy Journal*, 18: 63–83.

Oren, S., P. Spiller, P. Varaiya, and F. Wu (1995) "Nodal Prices and Transmission Rights: A Critical Appraisal," *The Electricity Journal*, 8(3): 24–35.

Perez-Arriaga, I. J. et. al (1995) "Marginal Pricing of Transmission Service: An Analysis of Cost Recovery," *IEEE Transactions on Power Systems*, 10(1): 546–553.

Stoft, S. (1999) "Financial Transmission Rights Meet Cournot: How TCCs Curb Market Power," *The Energy Journal*, 20: 1–23.

Vogelsang, I. (2001) "Price Regulation for Independent Transmission Companies," *Journal of Regulatory Economics*, 20(2): 141-165.

Williamson, O. (1983) "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, 73: 519–540.

Wilson, R. (2002) "Architecture of Power Markets," *Econometrica*, 70(4): 1299-1340.

Wu, F.P., Varaiya, P., Spiller, P., and S. Oren (1996) "Folk Theorems on Transmission Access : Proofs and Counterexamples," *Journal of Regulatory Economics*, 10(1): 5–24.