

IFN Working Paper No. 1351, 2020

Competition for Flexible Distribution Resources in a 'Smart' Electricity Distribution Network

Thomas P. Tangerås

Competition for flexible distribution resources in a 'smart' electricity distribution network*

Thomas P. Tangerås[†]

August 19, 2020

Abstract

In a 'smart' electricity distribution network, flexible distribution resources (FDRs) can be coordinated to improve efficiency. But coordination enables whoever controls such resources to exercise market power. The paper establishes the following efficiency rankings of market structures: Aggregators competing for FDRs are more efficient than a distribution system operator (DSO) controlling resources, which is more efficient than no FDR market. A no-market solution is more efficient than an FDR market featuring either (i) both DSO and aggregators; or (ii) a monopoly aggregator also supplying generation to the real-time market. The paper also characterizes a regulation that implements the efficient outcome.

Key words: Aggregator, distribution system operator, market power, real-time market, regulation, smart grid

JEL codes: H41; L12; L51; L94

*Many thanks to Therése Hindman Persson, Lennart Söder and seminar participants at the Swedish Competition Authority and the Swedish Energy Markets' Inspectorate for comments and discussion. This work was conducted within the "Sustainable Energy Transition" research program at IFN. Financial support from the Swedish Energy Agency (2015-002474) is gratefully acknowledged.

[†]Research Institute of Industrial Economics (IFN), PO. Box 55665, SE-10215 Stockholm, Sweden. Telephone: +46(0)86654526. E-mail: thomas.tangeras@ifn.se. Personal website: www.ifn.se/thomast. Associated researcher with the Energy Policy Research Group (EPRG), University of Cambridge. Faculty affiliate at the Program on Energy and Sustainable Development (PESD), Stanford University.

1 Introduction

A fundamental task of electricity system operation is to maintain a continuous balance between the electricity injected into the network and the electricity withdrawn from it.¹ Imbalances that are too large will cause electrical units and transmission lines to disconnect, which can develop into a serious outage unless contained.² In a restructured electricity market, the responsible system operator performs this task by procuring flexible reserve capacity through long-term mechanisms and/or in a real-time market.³ This capacity is activated in the degree necessary to offset real-time imbalances between consumption and production. Some system operators also possess reserve capacity of their own.

Over the last two decades, many countries have implemented support schemes to substantially increase the production of electricity from renewable energy sources. The trend towards more solar and wind power renders maintenance of system balance increasingly challenging. First, the demand for balancing power increases because the exogenous variability of solar and wind power makes it difficult to plan the output of those units in advance. Second, the supply of balancing power decreases because flexible thermal generation capacity, such as coal-fired plants and gas turbines, is competed out of the market by renewable units with zero or negative marginal production costs. The change in generation mix affects also the amount of hydro power available for balancing purposes because more hydro is needed to replace base-load thermal generation, such as nuclear power, forced out of the market.⁴ In sum, larger shares of intermittent renewable electricity generation imply that the market will supply less flexible generation capacity to cover an increased demand for balancing power.

The transformation of the electricity system has also started to affect the nature of system operation in the sense that imbalances have become more frequent and severe in the lower-voltage parts of the network. One explanation is that roof-top solar power and other energy resources connected to the distribution grid have become more widespread and render demand and supply at the local level increasingly difficult to predict. Phase-in of electric vehicles can be expected to add to future system complexity. The traditional approach to balancing the electricity system by expanding network capacity at all voltage levels and then leave it to a

¹The instantaneous balance is measured by the frequency with which the electric current oscillates through the grid. Network frequency increases (decreases) if the amount of electricity produced increases (decreases) relative to the amount that is consumed.

²All electrical units and machinery connected to the grid have the same nominal operating frequency. They will automatically disconnect if the actual frequency deviates too much from the nominal. The strain on the system caused by failed equipment can lead nearby units or network connections to become overloaded and also disconnect. In the worst case scenario, a domino effect ripples through the network and causes system collapse. A famous example is the Northeast blackout of 2003 that started with a power plant in Ohio shutting down. The failure spread through the system as transmission lines sequentially tripped offline. The ensuing outage affected some 10 million people in Ontario and 45 million people in eight U.S. states.

³Depending on the size of the market, there can be multiple system operators that each oversee one subset of the grid. In an international electricity market, the areas of responsibility are usually delineated by the national borders.

⁴Hydro and thermal power carry a substantial rotating mass in terms of spinning turbines. The associated rotational inertia contributes to system stability by absorbing instantaneous imbalances in production and consumption, much like a battery. The fact that solar and wind power have no or limited rotating mass increases the problem of maintaining stability by reducing system robustness.

centralized transmission system operator (TSO) to balance the entire system, is challenged when local variability in demand and supply is too large. This is particularly evident in those regions where network owners experience severe difficulties related to expansion of network capacity in the urban areas that need this capacity the most. Instead of continuing to rely entirely on centralized optimization at the transmission network level, some local imbalance problems might be solved more efficiently at the distribution network level (EDSO, 2015). In a more decentralized system, the responsibility for maintaining system balance within a local control area is delegated to a distribution system operator (DSO), most likely the owner of the local distribution network.

The advent of new technical solutions and improvements in information and communications technology have increased the potential for handling imbalance problems more at the local level. An increasing number of households install digital thermostats that collect data on heating and cooling. By matching these data with the prices of electricity, domestic heating systems are optimized to minimize heating costs. Similarly for plug-in vehicles, whose batteries are charged in such a way as to minimize the cost of using the vehicle. Micro generation can be combined with distributed battery technologies to minimize household electricity costs, as the price of batteries fall to a level at which they are commercially viable. Such *flexible distribution resources* (FDRs) can be used for a broader purpose than to optimize energy use for each individual household. The 'smart' electricity distribution network measures consumption and production at granular level and communicates with the different units in the network. Instead of having a strictly domestic application, thermostats and other devices can then improve overall system performance. Another example is the battery capacity in plug-in electric vehicles which can help balancing consumption and production in the grid more generally. In principle, the FDRs in the distribution network can be coordinated and reallocated in an effort to optimize the whole local electricity distribution system.⁵

This paper analyzes the emerging market for flexible distribution resources in relation to the utilization of these resources for system balancing purposes. A main problem is that control of substantial amounts of FDRs yields market power in the balancing market. The issue of market power raises a number of questions: How does the structure of the FDR market affect the efficiency with which FDRs are deployed? What are the consequences of allowing the DSO to participate in the FDR market? Is there a fundamental difference between a DSO controlling those resources compared to the case when the market for FDRs is supplied by one or more unbundled firms? What if a generation company supplies the FDR market? Is there a feasible way to regulate competition, and if so, what are the properties of such regulation? Answers to these questions can be helpful in making informed policy choices regarding competition and regulation of the evolving market for flexible distribution resources.

To analyze the performance of different market arrangements, I build in Section 2 a two-

⁵The term distributed energy resource (DER) is commonly used for describing a unit connected to the lower-voltage part of the network. Depending on the context, DERs may include also intermittent micro generation such as roof-top solar power and small-scale wind power. However, a DER must be *flexible* for it to be used for real-time balancing purposes. To emphasize this distinction, this paper introduces the concept of flexible distribution resources (FDRs), which are those DERs that can be dispatched in a controllable and flexible manner.

period model of an electricity system where demand fluctuates exogenously across periods. Maintaining system balance involves clearance of real-time consumption and production in each period. The standard way of balancing supply and demand is by adjusting flexible thermal generation within the control area or by importing electricity from or exporting electricity to neighboring control areas. I assume here that there is also a third way to accomplish system balance. Some of the electricity withdrawn from the grid goes into an energy service demanded by households. A prime example is indoor temperature control (heating and cooling). A constant indoor temperature can be achieved by different combinations of electricity consumption in the two periods. Household utility depends on keeping an ideal indoor temperature, not on how this temperature is produced. By these properties, the heating system is a flexible distribution resource because it effectively works as a battery that can be used for substituting electricity consumption across periods. Intertemporal substitution of household consumption reduces system cost because it offsets the variability in demand and thus stabilizes the production of balancing power. Section 2 shows that the system can be perfectly stabilized if the variation in demand is sufficiently small relative to the capacity for intertemporal substitution. If not, then it is first-best efficient to consume all electricity used in the provision of the energy service in the off-peak demand period. Section 2 also demonstrates how to implement this solution in a decentralized market if the supply of electricity is perfectly competitive and all market participants are exposed to marginal real-time prices.

There are plausible reasons why the decentralized solution would be infeasible even in a competitive real-time market for electricity.⁶ First, short-term pricing of electricity at most involves consumers being exposed to hourly, or perhaps half-hourly, price changes. Because of the real-time variability of renewable electricity generation, minimization of system costs should have consumers reacting to even more frequent price changes. Second, consumers with the ability or technology to respond to price changes even within a shorter time horizon may nevertheless dislike price variability and therefore be unwilling to take the price risk associated with implementing the efficient solution. Third, even if real-time prices were allowed to vary depending on instantaneous changes in underlying system conditions and consumers were indeed risk neutral, they could still have insufficient incentives to respond to them. This happens if the perceived household cost savings associated with real-time optimization are small relative to the costs, for instance because of incremental investments households would have to make to acquire the necessary technology. To account for such consumption inflexibility, I assume that households face average instead of granular real-time prices. Households then have no incentive to vary the consumption that goes into the production of the energy service. Instead, there are potential efficiency gains from third parties allocating FDRs across periods. Section 3 evaluates and compares different market solutions for such coordination.

I start the analysis of markets for flexible distribution resources by assuming in Section 3.1 that the DSO itself participates in the FDR market and does so without any competition. This is a natural starting point given that system operation originally was handled by a vertically

⁶Pollitt and Anaya (2019) discuss the potential for competition in short-term markets relative to the wholesale market.

integrated monopolist in the age before market restructuring. Households pay the DSO a fixed fee for supplying the energy service, but nothing for the electricity used in the production of the service. In return, the DSO assumes control over the FDR and manages the system by using the implicit storage capacity to allocate electricity consumption to the different periods. The first-best efficient solution is for the DSO to consume the larger share of the electricity used in the production of the energy service in the off-peak demand period up to the point of price equalization or exhaust the storage capacity, whichever comes first. This is *not* how a cost-minimizing DSO achieves system balance. Instead, it uses its market power to save on electricity costs by pushing down the price (or opportunity cost) in the off-peak relative to the peak demand period. The DSO accomplishes this price effect by increasing consumption in the peak and decreasing it in the off-peak demand period compared to the first-best efficient solution. The exercise of market power thus creates artificial differences in the cost of servicing electricity between the peak and off-peak period.

An increasing number of firms provide systems for household resource management. Examples include Converge, Energy Pool, Enbala and Enel X. Instead of simply minimizing household energy costs, such companies could equally well manage these resources by supplying balancing power to the real-time market. An example is the German company Sonnen that combines rooftop solar power with battery technology. Under the *sonnenFlat* plan, customers pay a fixed fee for an annual consumption allowance. Sonnen then takes control over the system and uses it to supply the household and provide balancing power by optimizing the use of the storage capacity in the system.⁷ A firm that coordinates distribution resources is known as an *aggregator*.⁸ In the current context, the aggregator performs such coordination to earn money in the local real-time market.⁹ If this is its only business, then the aggregator is *structurally independent*. Section 3.2 demonstrates that structurally independent aggregators competing in the market for flexible distribution resources increase efficiency by reducing real-time market power.

Section 3.3 considers the case where both the DSO and a number of structurally independent aggregators compete for customers in the market for FDRs. Regulatory policy may render it difficult for the DSO to prevent third-parties from supplying balancing power, but the DSO may nevertheless have discretion over the deployment of these resources across periods. The DSO will in this case allocate all aggregator consumption to the peak demand period and take advantage of the resulting price reduction in the off-peak period by withdrawing the electricity necessary to supply its own households in the off-peak period.

It makes a difference whether aggregators are structurally independent or integrated with generation units that supply electricity to the real-time market. By free entry into the market for flexible distribution services, it is likely that integration between supply and demand in the real-time market will occur. Section 3.4 shows that integration reinforces a monopoly aggregator's

⁷<https://sonnen.de/stromtarife/sonnenflat-home/>, October 15, 2019.

⁸See, for instance, Burger et al. (2017) for a conceptual discussion of aggregators and their potential role in the electricity system.

⁹Real-time markets at the level of the distribution network, sometimes known as *flexi markets*, are becoming increasingly common. For instance, the Nordic power exchange, Nord Pool, has developed the platform *Nodes* (<https://nodesmarket.com/>) for trading flexibility and energy at a decentralised level. One application of this platform is *sthlmflex* a regional flexibility market in the Stockholm area of Sweden.

incentive to increase the real-time price in the peak demand period because the firm then makes an additional profit on its flexible generation capacity.

Section 3.5 compares the different market structures in terms of system efficiency. A monopoly DSO allocates FDRs less efficiently than multiple structurally independent aggregators because of softer competition. Still, it is more efficient for the DSO to manage flexible distribution resources than not to coordinate them at all, in spite of these market power concerns, since the DSO always allocates more electricity to the off-peak demand period than in the benchmark case where consumption of electricity related to the energy service is completely inflexible. However, the distortions associated with a mixed market structure or an integrated monopoly aggregator are so severe that market efficiency is *lower* than in the case without any market for flexible distribution resources.

Section 3.6 illustrates how competition can be managed by way of compensation payments related to the supply of the flexible distribution resources in the real-time market, to ensure an efficient use of those resources. This regulation makes each firm a residual party to any efficiency loss it causes by its actions in the real-time market. Standard revenue regulation has no effect on the performance of the real-time market. The main effect of regulating revenue is to provide the regulated firms with a strong incentive to reduce costs. Exercise of market power in the real-time markets comes precisely from the incentive to minimize costs in the provision of the energy service.

Section 4 concludes with a discussion of implications for energy policy concerning markets for flexible distribution resources. Some proofs of formal statements are in the Appendix.

Contribution This is the first paper that examines how the incentives to deploy resources efficiently related to real-time balancing of consumption and production depend on the structure of the market for flexible distribution resources. Burger et al. (2019a,b) provide a comprehensive description of the core activities performed in an electricity distribution system that incorporates flexible distribution resources, and they characterize the actors as well as their potential roles in that system. Their policy analysis emphasizes governance structure in relation to competition and coordination in the distribution system, and in particular the incentives of a network owner that also controls FDR capacity to distort competition in the FDR market by restricting access to the network or to system services. The present paper complements this discussion, for instance by showing that distortions are likely to persist because of short-term market power even if the necessary investments have been made, and the regulatory framework can guarantee third-party equal access to the network.

Previous theoretical analysis has mainly focused on the incentives for network owners to procure such resources and how these incentives depend on regulatory policy; see for instance Brown and Sappington (2018, 2019). Kim et al. (2017) consider a model where multiple DSOs own local balancing capacity. Uncoordinated balancing is inefficient because of an assumed externality on the other DSOs. The paper devises an optimal cost-sharing mechanism of total balancing costs to ensure efficient short-term balancing of the market. In the present context, the single DSO does not own any capacity of its own. Instead, inefficiencies associated with local

short-term balancing of the market stem from the exercise of market power in the procurement and deployment of these services.

The modeling framework in this paper bears resemblance to the classical Hotelling model of resource extraction which considers the problem of how much of a finite resource to extract today and how much to save for the future. Under imperfect competition, the equilibrium in the Hotelling model is found at the point at which the marginal revenue is equalized across periods, barring any production constraints; see Tangerås and Mauritzen (2018) and the references therein. The present paper extends the Hotelling analysis by considering and comparing different market structures such as regulated monopoly, a mixed market structure and vertical integration between retail and production.

2 The first-best efficient benchmark

This section first presents the theoretical model and then characterizes the first-best efficient solution associated with centralized dispatch that will be used to evaluate the different market structures. I demonstrate that a real-time market design can implement this efficient solution if the market is perfectly competitive and all market participants face marginal real-time prices. I also identify a fundamental inefficiency associated with the standard market design in wholesale electricity markets where consumers are exposed to average instead of marginal prices.

The model Consider a two-period model of consumption and production within a control area. In each period $i = 1, 2$, there is exogenous and price-inelastic demand for $x_i > 0$ megawatt hours (MWh) electricity. This demand consists of industry and residential consumption minus the supply of intermittent and non-dispatchable renewable electricity, such as solar and wind power. I let the time frame between period 1 and 2 be so short that (x_1, x_2) is known at the start of period 1. This assumption and the below assumptions on cost functions simplify the analysis of the real-time market. There is also a representative household that consumes $\bar{s} > 0$ units of an energy service that is produced by withdrawing $s \geq 0$ MWh electricity from the grid in period 1 and $\bar{s} - s \geq 0$ MWh in period 2. One can think of the energy service as the production of heat (or cooling) to generate indoor temperature \bar{s} . Different combinations of s and $\bar{s} - s$ produce by the laws of thermodynamics a constant indoor temperature \bar{s} . The energy service effectively allows the household to substitute electricity consumption across periods. Linearity and perfect substitutability across periods are only to keep things simple.¹⁰

A fully flexible and dispatchable technology is available to cover total demand $q_1 = x_1 + s$ in period 1 and total demand $q_2 = x_2 + \bar{s} - s$ in period 2. It is most relevant to think of this technology as thermal electricity generation within the control area or as import and export capacity from and to neighboring control areas. Let $C(q_i)$ be the system cost of supplying the amount q_i of electricity to the control area in period $i = 1, 2$. The cost function $C(q)$ is the

¹⁰Intertemporal substitutability allows other interpretations of \bar{s} than the production of indoor temperature. In particular, one can view the energy service as a battery with capacity \bar{s} . Similarly, the energy service can be an electric vehicle that consumes electricity in amount \bar{s} .

same in both periods, continuous, differentiable and strictly increasing for all $q > 0$. The system marginal cost function $C'(q)$ is continuous, twice continuously differentiable, strictly increasing for $q > 0$, convex, and has bounded elasticity $\frac{C'''(q)q}{C''(q)} \leq 1$.¹¹

The first-best efficient allocation The central planner chooses first-period production $s \in [0, \bar{s}]$ of the household energy service to minimize the system total cost

$$C(x_1 + s) + C(x_2 + \bar{s} - s) \quad (1)$$

of clearing consumption and production in the two periods. Let s^{fb} be the amount of electricity used in the first period to produce the household energy service at the first-best efficient solution. Denote by $q_1^{fb} = x_1 + s^{fb}$ and $q_2^{fb} = x_2 + \bar{s} - s^{fb}$ production within the control area (or the net import from other control areas) in each of the two periods at the first-best efficient optimum.

Before characterizing the first-best efficient allocation, I introduce some terminology. The control area is *resource unconstrained* if fluctuations in demand across periods are small in the sense that $|x_1 - x_2| \leq \bar{s}$. Conversely, the control area is *resource constrained* if fluctuations in demand are so large so that $|x_1 - x_2| > \bar{s}$. Period 1 is the *peak demand period* and 2 the *off-peak demand period* if $x_1 > x_2$. The peak and off-peak definitions are reversed if $x_2 > x_1$.

Lemma 1 *Under the first-best efficient allocation $(s^{fb}, q_1^{fb}, q_2^{fb})$:*

(i) *The electricity used in the production of the household energy service is withdrawn from the grid in such a way as to smooth out all variations in system marginal costs across periods if the control area is resource unconstrained [$s^{fb} = s^* = \frac{1}{2}(\bar{s} + x_2 - x_1)$ and $q_1^{fb} = q_2^{fb} = q^* = \frac{1}{2}(\bar{s} + x_1 + x_2)$ if $|x_1 - x_2| \leq \bar{s}$].*

(ii) *The household energy service is produced entirely by withdrawing electricity from the grid in the off-peak demand period if the control area is resource constrained. In that case, local production (or the net import) is larger in the peak relative to the off-peak demand period [$s^{fb} = 0$ if $x_1 - x_2 > \bar{s}$ and $s^{fb} = \bar{s}$ if $x_2 - x_1 > \bar{s}$. Moreover, $(q_1^{fb} - q_2^{fb})(x_1 - x_2) > 0$ if $|x_1 - x_2| > \bar{s}$].*

The intertemporal aspect of the household production function allows to smooth out exogenous fluctuations in demand across periods and thus achieve full efficiency by eliminating the variability in the system marginal production costs, if the variability in renewable production is sufficiently small relative to the capacity for intertemporal substitution of electricity consumption, i.e. $|x_1 - x_2| \leq \bar{s}$. In more extreme cases, the central planner adapts to resource constraints by withdrawing all electricity that goes into producing the energy service in the off-peak demand period and clears excess demand by dispatching relatively more production resources or by increasing net imports in the period with the highest demand.

¹¹An example of a cost function with these properties is $C(q) = aq + bq^{2+\sigma}$, $a \geq 0$, $b > 0$, $\sigma \in [0, 1]$. The bounded elasticity assumption is used in Section 3.3 as a sufficient condition for comparative statics results and in Section 3.4 as a sufficient condition to ensure that the firm's profit function is strictly concave, but is otherwise redundant.

Decentralized market implementation I now establish conditions under which a decentralized market can implement the first-best efficient allocation. Assume that all production and consumption is cleared in a real-time market operating each period $i = 1, 2$. All production in period i is remunerated at the marginal real-time price $p_i = C'(q_i)$ and all consumption in that period also pays the marginal real-time price p_i .¹²

Assume that all market participants are price-takers. This implies that all thermal electricity or net import into the control area is bid in at marginal cost $C'(q_i)$. If all market participants expect the market to clear at the first-best efficient real-time price $p_i^{fb} = C'(q_i^{fb})$ in each period i , then the production of thermal electricity or net import into the control area equals q_i^{fb} in each period. Consumption of the energy service costs $p_1^{fb}s + p_2^{fb}(\bar{s} - s)$. If the control area is resource constrained because of peak demand in period 1, $x_1 - x_2 > \bar{s}$, then $p_1^{fb} = C'(q_1^{fb}) > C'(q_2^{fb}) = p_2^{fb}$ by Lemma 1. In this case, the household minimizes expenditures by consuming all electricity in the second period: $s = 0 = s^{fb}$. Conversely, the household consumes all electricity in the first period, $s = \bar{s} = s^{fb}$, if $x_2 - x_1 > \bar{s}$ because then $p_2^{fb} > p_1^{fb}$. Finally, the real-time price is the same in both periods, $p_1^{fb} = p_2^{fb} = p^* = C'(q^*)$, if the control area is resource unconstrained ($|x_1 - x_2| \leq \bar{s}$). In that case, the household's total expenditure $p^*\bar{s}$ on the energy service is independent of s , so it is individually rational to set $s = s^* = s^{fb}$. Hence, the competitive real-time market implements the first-best efficient allocation as a decentralized equilibrium. The following implications are straightforward:

Corollary 1 *In a competitive real-time market that implements the first-best efficient allocation $(s^{fb}, q_1^{fb}, q_2^{fb})$, the real-time price is weakly higher in the peak than the off-peak period and the representative household consumes relatively more electricity in the low-price compared to the high-price period $[(p_1^{fb} - p_2^{fb})(x_1 - x_2) \geq 0$ and $(p_1^{fb} - p_2^{fb})(\frac{1}{2}\bar{s} - s^{fb}) \geq 0$ with strict inequalities if and only if $|x_1 - x_2| > \bar{s}]$.*

The cost of inflexibility As argued in the introduction, there are plausible reasons why the decentralized solution would be infeasible even in a competitive market. In particular, households typically are not exposed to marginal real-time prices. Instead, consumption decisions are most often based on average prices. To account for this lack of contractual flexibility, I henceforth assume that representative household pays the average real-time price $\frac{1}{2}(p_1 + p_2)$ for consumption of the energy service. Households then cannot strictly benefit from varying s across the two periods. Instead, it is optimal for them to withdraw the same amount of electricity from the grid in period 1 as in period 2 to produce the household energy service, in which case $s = \frac{1}{2}\bar{s}$. Then the thermal electricity production in the control area or net import from neighboring areas that solves the real-time balancing problem in period $i = 1, 2$ equals $\bar{q}_i = x_i + \frac{1}{2}\bar{s}$. This solution generally is inefficient:

Proposition 1 *A representative household exposed to the average real-time price consumes too much electricity in the peak relative to the off-peak demand period. Real-time system balancing*

¹²An alternative market design could be pay-as-bid pricing.

requires excessive local production (or net import) in the peak demand period to offset this inefficiency [$(\frac{1}{2}\bar{s} - s^{fb})(x_1 - x_2) > 0$, $(\bar{q}_1 - q_1^{fb})(x_1 - x_2) > 0$ and $(\bar{q}_2 - q_2^{fb})(x_2 - x_1) > 0$ for all $x_1 \neq x_2$]

Proof: By way of Lemma 1, $(\frac{1}{2}\bar{s} - s^{fb})(x_1 - x_2) = \frac{1}{2}(x_1 - x_2)^2$ for $|x_1 - x_2| \leq \bar{s}$ and $(\frac{1}{2}\bar{s} - s^{fb})(x_1 - x_2) = \frac{1}{2}\bar{s}|x_1 - x_2|$ for $|x_1 - x_2| > \bar{s}$. To see the second part of the proposition, plug in (q_1^{fb}, q_2^{fb}) and (\bar{q}_1, \bar{q}_2) to get $(\bar{q}_1 - q_1^{fb})(x_1 - x_2) = (\bar{q}_2 - q_2^{fb})(x_2 - x_1) = \frac{1}{2}(x_1 - x_2)^2$ for $|x_1 - x_2| \leq \bar{s}$ and $(\bar{q}_1 - q_1^{fb})(x_1 - x_2) = (\bar{q}_2 - q_2^{fb})(x_2 - x_1) = \frac{1}{2}\bar{s}|x_1 - x_2|$ for $|x_1 - x_2| > \bar{s}$. ■

The $(\frac{1}{2}\bar{s}, \bar{q}_1, \bar{q}_2)$ allocation is inefficient relative to the first-best allocation $(s^{fb}, q_1^{fb}, q_2^{fb})$ except in the knife-edge case where exogenous demand is the same in both periods, $x_1 = x_2$, because the difference in system marginal costs across periods is no longer minimized. This inefficiency opens up for complementary solutions to a fully decentralized market with complete real-time price exposure to correct these distortions. This paper considers a market for flexible distribution resources (FDR). In the specific context of the present model, one can think of FDR as the intertemporal production of the household energy service.

3 A market for flexible distribution resources

Assume from now on that neither the centralized solution nor the fully decentralized market with complete real-time price exposure for all consumers is feasible. Instead, a distribution system operator (DSO) is responsible for maintaining system stability within the control area in each period. This DSO has no production capacity of its own and therefore has to procure the required balancing power. There are two ways in which the DSO can do this. First, the DSO can import electricity from or export electricity to surrounding price control areas at the going short-term market price and/or sign contracts with local generation capacity for balancing power. Second, the DSO can implement a local real-time market. Either way, I assume that thermal generation within the control area (or net traded electricity with surrounding areas) is supplied at marginal cost and that all activated balancing power receives the same compensation, equal to the marginal cost of the most expensive unit that is activated. Therefore, $P(q_i) = C'(q_i)$ represents the inverse supply function of real-time electricity q_i in period i . The assumption that thermal electricity is supplied by generation owners at marginal cost allows to isolate the effects on efficiency of imperfect competition in the market for flexible distribution resources. Yet, many of the results would carry over to a setting with market power also in the supply of thermal electricity; see the discussion in Section 4.

Smart grid solutions can be utilized in order to accomplish system balance. Specifically, the network infrastructure is 'smart' in the sense that an external party can assume the task of supplying the energy service \bar{s} to the representative household by remote control of s . The economic incentive to assume this responsibility comes from the ability to supply the flexible distribution resource to balance electricity supply within the local control area. I refer to \bar{s} as the size of the market for flexible distribution resources (FDRs). The key policy question is how this market should be structured to maximize efficiency. To address this question, I

examine four market structures. In Section 3.1, the DSO has monopoly in the FDR market. In Section 3.2, $A \geq 1$ structurally independent aggregators compete in the FDR market. Section 3.3 analyzes the properties of a mixed market structure in which both the DSO and $A \geq 1$ structurally independent aggregators compete in the FDR market. Section 3.4 considers the case of an aggregator that also owns generation capacity that supplies balancing power. In Section 3.5, I compare the efficiency properties of the different market structures. I then characterize price regulation of flexible distribution resources to implement the first-best efficient outcome in Section 3.6.

3.1 Monopoly DSO

I assume that the DSO is subject to revenue regulation by way of its monopoly position in the grid. This implies that there is a cap on the amount of revenues the DSO can collect from customers. Often, the cost $P(q_1)q_1 + P(q_2)q_2$ of procuring balancing power is considered unavoidable and therefore added in full to the revenue cap. However, the DSO would have no incentive to incorporate cost-reducing FDRs into system operation under full pass-through of balancing costs, but a strict incentive *not* to do so if implementation of FDR solutions involves non-monetary costs the DSO cannot pass onto consumers (Kim et al., 2017). A second possibility, and the one we explore in this section, is to add only a fraction $1 - \theta \in [0, 1)$ of the energy service costs $P(q_1)s + P(q_2)(\bar{s} - s)$ to the revenue cap and require the DSO to pay the remaining share θ out of its own pocket. The DSO then has an incentive to introduce FDR solutions for θ sufficiently large. A third possibility would be to apply specific regulation of the FDR market, a topic we return to in Section 3.6.

The market for flexible distribution resources operates as follows in this particular setting. At the outset, the DSO approaches the representative household and offers to supply the energy service \bar{s} in return for a fixed fee t . The DSO cannot charge a fee that exceeds the household's expected cost $\frac{1}{2}(p_1 + p_2)\bar{s}$ of purchasing the electricity for the household service. However, as the fixed fee enters into the revenue cap, it does not cost anything for the DSO to satisfy the representative household's participation constraint as it can simply increase other network tariffs correspondingly.

In the second stage, the DSO delivers the energy service by withdrawing s from the grid in period 1 and $\bar{s} - s$ in period 2. The profit of the DSO equals

$$\Pi^{DSO}(s) = F - \theta[P(x_1 + s)s + P(x_2 + \bar{s} - s)(\bar{s} - s)] \quad (2)$$

because it must cover a fraction θ of the total energy service costs out of its own pocket. In the above equation, F is the revenue cap, which we treat as exogenous. Two things are worth noticing about the profit expression $\Pi^{DSO}(s)$. First, electricity withdrawal s matters for profit even if the DSO is regulated because the DSO is residual claimant to a fraction of associated savings. Second, the DSO faces the *inverse supply function* $P(q_i)$ in the real-time market in period i . In other words, a DSO with monopoly control over flexible distribution resources can exercise monopsony power. This is unlike in the benchmark case of a fully decentralized market

where s is decided by a representative household that takes prices as exogenously given.

The DSO has an incentive to allocate consumption to the first period (increase s) if $p_1 < p_2$:

$$\Pi^{DSO'}(s) = \theta[p_2 - p_1 - P'(q_1)s + P'(q_2)(\bar{s} - s)]. \quad (3)$$

But it also takes into account the increase in the first period marginal system cost and the reduction in the second period marginal system cost resulting from the reallocation. This market power effect is identified by the two last terms in the marginal profit expression (3). Observe also that the trade-off is independent of θ for all $\theta > 0$. Let $s^{DSO} \in [0, \bar{s}]$ be the equilibrium amount of electricity withdrawn by the DSO in period 1 to supply the energy service to households, so that $\bar{s} - s^{DSO}$ is the amount withdrawn in period 2. Then $q_1^{DSO} = x_1 + s^{DSO}$ and $q_2^{DSO} = x_2 + \bar{s} - s^{DSO}$ measure the equilibrium amount of electricity supplied to balance the market in period 1 and 2, respectively.

Proposition 2 *Consider the equilibrium allocation $(s^{DSO}, q_1^{DSO}, q_2^{DSO})$ under the assumption that the DSO is residual claimant to a fraction $\theta \in (0, 1]$ of energy service cost savings and has a monopoly position in the market for flexible distribution resources:*

- (i) *Electricity withdrawal is upward-distorted in the peak demand period compared to the first-best efficient allocation $[(s^{DSO} - s^{fb})(x_1 - x_2) \geq 0$ with strict inequality if $|x_1 - x_2| \in (0, \bar{s})]$.*
- (ii) *The DSO withdraws more electricity from the grid in the off-peak relative to the peak demand period to supply the household energy service $[(\frac{1}{2}\bar{s} - s^{DSO})(x_1 - x_2) > 0$ for all $x_1 \neq x_2]$.*
- (iii) *The equilibrium allocation is independent of cost-sharing θ .*

The DSO wields market power by its control of \bar{s} . Its objective is to minimize the total expenditures associated with supplying the energy service, not to minimize total system cost. To see how market power affects the equilibrium allocation, consider a classical "bathtub" diagram depicted in Figure 1 below. The horizontal axis measures the withdrawal of electricity in the first period from left to right and in the second period from right to left. At $s = 0$, all electricity used in the production of the household energy service is consumed in period 2, and at $s = \bar{s}$ it is consumed in its entirety in period 1. The left-most vertical axis measures marginal system costs and marginal expenditures in period 1 and the right-most axis the marginal system costs and marginal expenditures in period 2. The two curves $P(q_1)$ and $P(q_2)$ are the inverse short-term supply functions of electricity $q_1 = x_1 + s$ in period 1 and $q_2 = x_2 + \bar{s} - s$ period 2. At $s = \frac{1}{2}\bar{s}$, the system marginal cost is smaller in period 1 than period 2, $P(x_1 + \frac{1}{2}\bar{s}) < P(x_2 + \frac{1}{2}\bar{s})$, thus establishing period 2 as the peak-demand period, $x_2 > x_1$.

An increase in electricity withdrawal s in period 1 drives up the system marginal cost in period 1, but reduces it in period 2. The control area is resource unconstrained, so the first-best efficient allocation is found at the point $s = s^{fb} \in (\frac{1}{2}\bar{s}, \bar{s})$ at which the system marginal costs are equated across the two periods: $P(x_1 + s^{fb}) = P(x_2 + \bar{s} - s^{fb}) = p^{fb}$. However, the marginal benefit to the DSO of increasing s is not measured in terms of the difference $P(q_2) - P(q_1)$ in system marginal costs. Instead, the DSO accounts also for the increase in the system marginal cost in period 1 and the decrease in period 2 because these changes affect the total payments

to generators and compensation for net imports from other control areas in each period. The marginal expenditure of increasing electricity withdrawal in period 1 is illustrated in Figure 1 from left to right by $P(q_1) + P'(q_1)s$, whereas the marginal expenditure of increasing electricity withdrawal in period 2 is illustrated the figure from right to left to right by $P(q_2) + P'(q_2)(\bar{s} - s)$. Starting at the first-best efficient allocation s^{fb} , a marginal reduction in s below s^{fb} reduces the period 1 expenditure by $p^{fb} + P'(q_1)s^{fb}$ and increases it by $p^{fb} + P'(q_1)(\bar{s} - s^{fb})$ in period 2 because the DSO reallocates electricity consumption from the off-peak demand period 1 to the peak demand period 2. This manipulation is strictly profitable to the DSO because electricity withdrawal is strictly larger in period 1 than period 2, $s^{fb} > \frac{1}{2}\bar{s}$. The allocation that minimizes the DSO's total expenditures is found at the point $s^{DSO} \in (\frac{1}{2}\bar{s}, s^{fb})$ at which the marginal expenditure is the same in both periods.

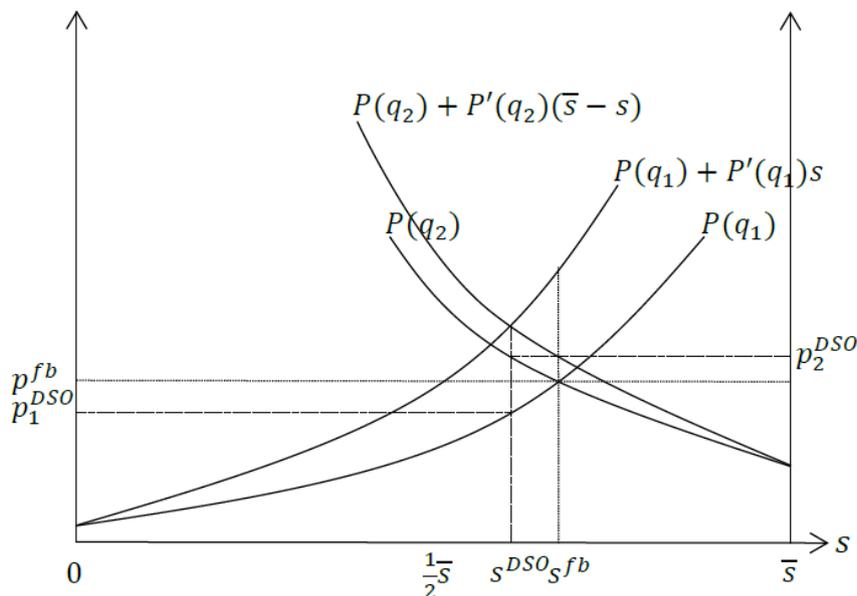


Figure 1: Monopoly DSO

Exploitation of market power causes the DSO to withdraw too much electricity from the grid in the peak period and too little in the off-peak period. However, the DSO still withdraws more electricity from the grid in the off-peak than the peak demand period because the difference in demand implies that the marginal cost is lower in the off-peak compared to the peak demand period, and the DSO therefore can reduce spending by allocating relatively more consumption to the off-peak demand period.

The exercise of market power in Figure 1 drives up the marginal cost of electricity in period 2 relative to period 1 in equilibrium, $p_2^{DSO} > p_1^{DSO}$, even if there is no real scarcity of resources, $|x_1 - x_2| \in (0, \bar{s}]$, so that full equalization of marginal costs would be feasible and efficient. The following corollary to Proposition 2 therefore arises:

Corollary 2 *Monopoly control over the flexible distribution resource by the DSO and exercise*

of market power in the market for balancing power lead to excessive fluctuations in marginal costs by which the DSO inflates [deflates] the marginal cost in the peak [off-peak] demand period $[(p_1^{DSO} - p_2^{DSO})(x_1 - x_2) > 0 = (p_1^{fb} - p_2^{fb})(x_1 - x_2)]$ for $|x_1 - x_2| \in (0, \bar{s}]$.

Market power reduces the value of letting the DSO supply the energy service because of the associated distortions to resource allocation, Still, this market structure is not obviously a bad idea in this model, in particular if the inflexible allocation $(\frac{1}{2}\bar{s}, \bar{q}_1, \bar{q}_2)$ represents the default situation. The DSO still allocates more electricity to the off-peak period compared to the outside option. Under sufficient variability of renewable electricity, the DSO even supplies the energy service efficiently, $s^{DSO} = s^{fb} \in \{0, \bar{s}\}$, despite its market power.

The common regulatory policy of regulating DSOs by way of a revenue cap leads to an exercise of monopsony power by the DSO which is exactly the same as if DSO operations in the FDR market were completely unregulated. As an alternative to more detailed regulation, it is therefore interesting to consider other market solutions than DSO monopoly supply of flexible distribution resources.

3.2 Structurally independent aggregators

Assume that the DSO clears imbalances by way of a real-time (flexi) market and that its only task is to operate this market. In particular, the DSO does not participate in the market for flexible distribution resources. All financial transactions in the real-time market are budget-balanced. A set of structurally independent aggregators compete in the market for flexible distribution resources. These are profit-maximizing firms whose only business in this model is to supply the energy service \bar{s} to households. These aggregators simultaneously and independently compete for a continuum of representative households with measure one and then purchase electricity in the real-time market to fulfill their supply obligations.

Let aggregator $a = \{1, \dots, A\}$, $A \geq 1$, charge a fee $t_a \geq 0$ in the first stage, and assume that it thereby obtains a market share $L_a > 0$, $\sum_{a=1}^A L_a = 1$, in the second stage. Aggregator a then purchases $s_a \in [0, L_a \bar{s}]$ in the real-time market in period 1 and the remaining $L_a \bar{s} - s_a$ in the second period to supply energy services in total quantity $L_a \bar{s}$. Let $\sum_{a=1}^A s_a = s$ be the A aggregators' total demand for real-time electricity in period 1. Then $\bar{s} - s$ is their total demand in period 2. Under the assumption that aggregators compete in quantities in the real-time market, the profit of aggregator a is

$$\Pi^A(s_a) = L_a t_a - P(x_1 + s) s_a - P(x_2 + \bar{s} - s)(L_a \bar{s} - s_a). \quad (4)$$

Under symmetry, $L_a = \frac{1}{A}$ and $s_a = \frac{1}{A} s$, the marginal incentive to increase energy withdrawal in the first period (increase s_a) equals:

$$\Pi^A(\frac{1}{A} s) = p_2 - p_1 - \frac{1}{A} P'(q_1) s + \frac{1}{A} P'(q_2) (\bar{s} - s). \quad (5)$$

Aggregators have an incentive to allocate electricity purchases to the period with the lowest real-time price. However, market power causes aggregators to behave differently than the central

planner because aggregators also take into account how their demand in the real-time market affects real-time prices. Yet the marginal price effects in (5) are smaller in magnitude than those for the DSO in (3) if $A \geq 2$ because some of the marginal price effects then spill over to the other aggregators in the market.

Let $s^A \in [0, \bar{s}]$ be the total amount of electricity withdrawn from the grid by the A aggregators in period 1 and $\bar{s} - s^A$ the total amount of electricity withdrawn from the grid in period 2 to supply the energy service in symmetric equilibrium where $L_a = \frac{1}{A}$ and $s_a^A = \frac{1}{A}s^A$. Then $q_1^A = x_1 + s^A$ and $q_2^A = x_2 + \bar{s} - s^A$ measure the equilibrium supply in period 1 and 2, respectively, in the market with aggregators.

Proposition 3 *Consider the symmetric equilibrium allocation (s^A, q_1^A, q_2^A) under the assumption that the market for flexible distribution resources consists of $A \geq 1$ structurally independent and symmetric aggregators competing for customers:*

(i) *Electricity withdrawal is (weakly) upward-distorted in the peak demand period compared to the first-best efficient allocation, but less so when there are more aggregators $[(s^A - s^{fb})(x_1 - x_2) \geq 0$ with strict inequality if $|x_1 - x_2| \in (0, \bar{s})$. Moreover, $\frac{d}{dA}(s^A - s^{fb})(x_1 - x_2) < 0$ if $s^A \in (0, \bar{s})$ and $x_1 \neq x_2$].*

(ii) *Aggregators withdraw more electricity from the grid in the off-peak and less electricity in the peak demand period than a DSO with monopoly power $[(s^{DSO} - s^A)(x_1 - x_2) \geq 0$ with strict inequality if and only if $A > 1$, $x_1 \neq x_2$ and $s^{DSO} \in (0, \bar{s})$].*

Aggregators exercise market power by consuming too much electricity in the peak demand period, but the fragmented market structure mutes the incentive to exercise market power. By implication, real-time prices are more stable in a market with independent aggregators compared to the monopoly DSO solution.

3.3 Mixed market structure

An important policy question in the evolving market for flexible distribution resources is whether the DSO itself should be allowed to participate in that market. The case analyzed in Section 3.1, where only the DSO is capable of supplying this service, offers some insights. Consider now the case where the DSO and $A \geq 1$ structurally independent aggregators compete for flexible distribution resources. Assume that the DSO obtains $\bar{s}^{DSO} > 0$ of the total market $\bar{s} > 0$ for flexible distribution resources by charging the fixed tariff t^{DSO} for supplying the energy service. The A aggregators have the rest of the market: $\bar{s}^A = \bar{s} - \bar{s}^{DSO} > 0$.

A classical concern in electricity markets is that a network owner with commercial interests can benefit from limiting competing commercial interests' access to the network. Such direct foreclosure (Rey and Tirole, 2007) is perhaps not a major problem under a standard regulatory policy, which in the current setting mandates the DSO to grant aggregators access to the network on non-discriminatory terms. However, this equal network access will turn out to be insufficient because the DSO can hurt aggregators also through its system operations. This possibility is particularly obvious if aggregators do not participate directly in a local real-time market,

but instead pay the cost of their real-time electricity consumption to the DSO. Such a market design will occur, for instance, if the DSO does not operate any formal real-time market. In this case, the DSO can drive aggregators out of the FDR market by overcharging them for system operation costs. Full pass-through to consumers of system operation payments collected from aggregators will not solve this problem. A solution to the problem of foreclosure through inflated system costs is instead to establish a formal real-time market in which the aggregators can purchase their electricity. The costs of the aggregators' electricity consumption then depend on the market-clearing real-time prices.

Assume that the DSO operates a formal real-time market as described in Section 2. The DSO profit

$$\Pi(s^{DSO}, s^A, x_1, x_2) = \frac{\bar{s}^{DSO}}{s} t^{DSO} - P(x_1 + s^{DSO} + s^A) s^{DSO} - P(x_2 + \bar{s} - s^{DSO} - s^A) (\bar{s}^{DSO} - s^{DSO}) \quad (6)$$

(assuming $\theta = 1$, so that the DSO is residual claimant to all cost savings) depends on the amount of electricity $s^{DSO} \in [0, \bar{s}^{DSO}]$ it withdraws in the first period [and $\bar{s}^{DSO} - s^{DSO}$ in period 2] to supply the energy service to the households served by the DSO itself, and on the amount $s^A \in [0, \bar{s}^A]$ of electricity purchased by the A aggregators in the real-time market in the first period [and $\bar{s}^A - s^A$ in period 2] to supply the energy service to households that purchase this service from an aggregator. Then, $q_1 = x_1 + s^{DSO} + s^A$ is the supply of electricity in period 1, and $q_2 = x_2 + \bar{s} - s^{DSO} - s^A$ is the supply in period 2. The corresponding real-time price in period $i = 1, 2$ is given by $p_i = P(q_i)$. If demand (x_1, x_2) and the marginal cost function $C(q)$ are common knowledge, then the equilibrium of the mixed market structure is the same as the one with $A + 1$ structurally independent aggregators. In this case, competition increases by allowing the DSO to participate in the real-time market.

Two necessary conditions of the above solution are likely to be violated in an actual real-time market. The first is the assumption of fully decentralized dispatch. In reality, participants in a real-time market commit their capacity for a fixed time interval. The system operator then dispatches electricity to accomplish an instantaneous balance of production and consumption within the duration of the time interval. A simple way to incorporate this feature into the model is to assume that suppliers commit to the inverse supply function $C(q)$ for both periods before the start of period 1, whereas aggregators commit to purchasing \bar{s}^A MWh electricity over the two periods as a whole. The actual dispatch of this capacity in period 1 and 2 is left to the DSO. Under these assumptions, the DSO chooses (s^{DSO}, s^A) to maximize profit (6).

The second condition is the complete information assumption. In reality, the DSO is likely to possess superior information about the real-time market simply because the DSO is the one that collects the bids and offers from the market participants and operates the system. The DSO can use this private information to its advantage by distorting the information about supply and demand conditions released to the other market participants. To isolate the implications of DSO private information about system operation, let us maintain the assumption of decentralized dispatch in each period. Suppose the DSO can manipulate the real-time market by exposing aggregators to a different supply function $\hat{P}(q_i)$ in period i than the actual $P(q_i)$ submitted by

suppliers. The DSO can effectively implement any (s^A, s^{DSO}) it desires also in this setting. To see this, notice that the profit of representative aggregator a is

$$\frac{1}{A} \frac{\bar{s}^A}{\bar{s}} t^A - \hat{P}(x_1 + s^{DSO} + \frac{A-1}{A} s^A + s_a) s_a - \hat{P}(x_2 + \bar{s} - s^{DSO} - \frac{A-1}{A} s^A - s_a) (\frac{1}{A} \bar{s}^A - s_a)$$

in the market with manipulated supply functions, under the simplifying assumption that all aggregators charge the same tariff t^A , have the same market share, and all aggregators other than a purchase $\frac{1}{A} s^A$ in the real-time market in period 1 and $\frac{1}{A} (\bar{s}^A - s^A)$ in period 2. The marginal profit of aggregator a reads

$$\hat{P}(q_2) - \hat{P}(q_1) - \frac{1}{A} \hat{P}'(q_1) s^A + \frac{1}{A} \hat{P}'(q_2) (\bar{s}^A - s^A)$$

in the equilibrium with symmetric aggregator demand, $s^a = \frac{1}{A} s^A$. By setting $\hat{P}(q_i) = p_i$ for $i = 1, 2$, the DSO ensures budget balance in the real-time market in each period i because all real-time electricity is traded at the uniform price p_i . The DSO then implements its preferred s^A by manipulating the slope $\hat{P}'(q_1)$ and/or $\hat{P}'(q_2)$. Hence, the DSO incites aggregators into purchasing s^A in the first period and $\bar{s}^A - s^A$ in the second by exposing them to *perturbed elasticities of the residual supply-curves*, not by perturbing real-time prices.

Consider now the DSO's profit-maximizing choice (s^{DSO}, s^A) in the mixed market structure. Let period 2 be the peak demand period, $x_2 > x_1$. Starting at the first-best efficient solution s^* , the DSO has an incentive to increase the peak price p_2 and reduce the off-peak price p_1 to save on costs; see the discussion in Section 3.1. The way to accomplish this price change in the DSO monopoly case is by reducing the withdrawal of electricity s^{DSO} in off-peak period 1 relative to peak period 2. In the mixed market structure, the DSO can accomplish this price manipulation in a much more profitable way by instead reducing *aggregator* consumption s^A in off-peak period 1. At the corner solution, $s^{DSO} = \bar{s}^{DSO}$ and $s^A = 0$, the DSO reaps the full benefit of the lower real-time price in off-peak period 1 whereas the full burden of the higher real-time price in peak period 2 falls upon the aggregators.

Proposition 4 *In a mixed structure where the DSO and $A \geq 1$ structurally independent aggregators compete in the market for flexible distribution resources:*

- (i) *Aggregators consume all their electricity in the peak demand period [$s^A = \bar{s}^A$ if $x_1 > x_2$, $s^A = 0$ if $x_1 < x_2$].*
- (ii) *More electricity is withdrawn from the grid in the peak relative to the off-peak demand period to produce the household energy service if the DSO controls half or less of the market for flexible distribution resources [$(s^{DSO} + s^A - \frac{1}{2} \bar{s})(x_1 - x_2) \geq 0$ if $\bar{s}^{DSO} \leq \frac{1}{2} \bar{s}$].*

To ensure that the real-time market works as efficiently as possible if the DSO also participates in the market for flexible distribution resources, it is important to ensure that the DSO is unable to manipulate the real-time market by releasing distorted information about bids and offers. One way is to require of participants in the real-time market that they simultaneously submit their bids to the DSO and a third party, for instance a regulatory authority. Such information sharing enables the authority to verify ex post that the information released to

market participants did not differ substantially from the actual bids. Preventing the DSO from exercising market power associated with centralized dispatch of FDRs is considerably more challenging. Performing the appropriate counterfactual analysis requires exact information about all relevant aspects of the historical operating conditions. This task is exacerbated by the fact that a multitude of unit-specific properties, such as ramping constraints and location in the network, determine the efficient real-time dispatch of resources. If it is not possible to separate the task of balancing production and consumption from the DSOs other tasks, such as the supply of flexible distribution resources, the optimal policy recommendation could be to prohibit the DSO from participating in the FDR market.

3.4 Integrated aggregator

It is likely that also incumbent firms already present in the balancing market and not only structurally independent aggregators, will compete in the market for flexible distribution resources. We consider here an incumbent firm that is integrated between retail and production in the sense of participating on both sides of the balancing market. To consider the effects of such integration, assume that only one firm participates in the market for FDRs and that this firm also owns all thermal generation capacity within the control area. The fixed fee t^I this integrated aggregator charges for supplying the energy service matters for income redistribution, but not for efficiency. I maintain the assumption that thermal electricity is competitively supplied so that the period i inverse supply function equals $p_i = P(q_i) = C'(q_i)$. This assumption facilitates the comparison with previous results, but I discuss the consequences of market power in thermal generation below. The profit of the integrated aggregator then equals as a function of s :

$$\Pi^I(s) = t^I + P(x_1 + s)x_1 + P(x_2 + \bar{s} - s)x_2 - C(x_1 + s) - C(x_2 + \bar{s} - s). \quad (7)$$

Let $s^I \in [0, \bar{s}]$ be the equilibrium amount of electricity withdrawn from the grid by the integrated aggregator in period 1 and $\bar{s} - s^I$ the equilibrium amount of electricity withdrawn from the grid in period 2 to supply the energy service. Let $q_1^I = x_1 + s^I$ and $q_2^I = x_2 + \bar{s} - s^I$ measure the equilibrium amount of electricity supplied in period 1 and 2, respectively, in the market with one integrated aggregator.

Proposition 5 *An integrated aggregator that bids its thermal generation competitively into the real-time market and has a monopoly in the market for flexible distribution resources, withdraws more electricity from the grid in the peak relative to the off-peak demand period $[(s^I - \frac{1}{2}\bar{s})(x_1 - x_2) \geq 0$ with strict inequality if $x_1 \neq x_2$].*

The distortions to the real-time market arising from the exercise of market power in the supply of the energy service \bar{s} are exacerbated if the service is provided by a firm that also supplies thermal electricity to the market, compared to the case of a structurally independent monopoly aggregator. This aggregator has an excessive incentive to reallocate electricity consumption to the peak demand period because the associated price reduction in the off-peak period generates cost savings that dominate the cost increase resulting from a price increase in the peak demand

period. Compared to the independent aggregator, the integrated aggregator has an even stronger incentive to allocate electricity consumption to the peak demand period because it can then increase the income on the generation capacity it bids into the market. If this generation capacity is bid in competitively, then the incentive to allocate electricity consumption to the peak period is so strong that it causes the integrated aggregator to withdraw more electricity from the grid in the peak than the off-peak demand period.

The fundamental difference between a producer selling energy services to households and an independent aggregator doing the same, is that the former market structure involves integration between production and retail. These results establish how such vertical integration enables a firm to exercise market power by bidding up prices on the demand side of the market. The net effect is a reduction in overall market efficiency if the market is otherwise competitive. However, it is well-known that integration can improve competition on the production side (e.g. Wolak, 2007 and Bushnell et al., 2008). Thermal producers with market power withhold output from the market in both periods to increase the real-time price. Integration between production and retail implies that there is a benefit to increasing thermal output because lower real-time prices reduce the cost of providing the energy service. Hence, with imperfect competition both on the production and the retail side, the net effect on efficiency of generation and retail integration is less clear-cut. The overall conclusions are likely to depend on the elasticity of demand across periods relative to the elasticity of demand in the real-time market, but I leave this issue for future research.

3.5 Comparison of market structures

The efficiency with which flexible distribution resources are allocated to the real-time market depends fundamentally on the structure of the market for flexible distribution resources:

Proposition 6 *The different structures of the market for flexible distribution resources can be ranked in decreasing order of system cost efficiency:*

- (i) $A \geq 1$ structurally independent and symmetric aggregators.
- (ii) Monopoly DSO.
- (iii) No FDR market (e.g. households subject to average real-time prices).
- (iv) Mixed market structure (with $\bar{s}^{DSO} \leq \frac{1}{2}\bar{s}$); or
- (v) Integrated monopoly aggregator with competitive supply of thermal electricity.

Proof: By way of propositions 2, 3, 4 and 5, it follows that $s^{fb} \leq s^A \leq s^{DSO} < \frac{1}{2}\bar{s} \leq \min\{s^{DSO} + s^A; s^I\}$ if $x_1 > x_2$ and $\max\{s^{DSO} + s^A; s^I\} \leq \frac{1}{2}\bar{s} < s^{DSO} \leq s^A \leq s^{fb}$ if $x_2 > x_1$. This ranking of the equilibrium s under the different market structures and the strict convexity of the system total cost $C(x_1 + s) + C(x_2 + \bar{s} - s)$ in s yield the result. ■

The real-time market is exposed to firms' exploitation of market power to a larger or smaller extent depending on the structure of the market for flexible distribution resources. Notwithstanding market power issues, and for relevant market structures, it is better to introduce a market for flexible distribution resources than to maintain a market design that does not utilize

this resource at all. Important exceptions to this rule occur when the DSO and aggregators are simultaneously present in the FDR market, or when an aggregator is active on both sides of the real-time market by also owning generation capacity. These findings suggest that sufficient economic separation is important for the desirability of a market for flexible distribution resources.

3.6 Efficient regulation

Unfettered competition for flexible distribution resources generally leads to inefficient market allocations if this market is imperfectly competitive. An option then is to introduce some type of regulation if one cannot establish well-functioning competition, for example because of entry barriers. This regulation will have to be more fine-grained than the one currently imposed on distribution network owners. In particular, standard revenue cap regulation will have no effect on market power compared to the situation when firms are not regulated at all; see Section 3.1.

To illustrate the type of regulatory scheme that will implement a first-best efficient outcome in this setting, consider the case of a structurally independent aggregator that holds a monopoly position in the market for flexible distribution resources, i.e. $A = 1$ in Section 3.2. Assume that this aggregator sells the energy service to the representative household at the fixed price t^R . The outside option of the household is to purchase the electricity needed for production of the energy service at expected cost $\frac{1}{2}(p_1^{fb} + p_2^{fb})\bar{s}$ if the household expects first-best efficient real-time prices. Then $t^R = \frac{1}{2}(p_1^{fb} + p_2^{fb})\bar{s}$ is the maximal tariff the aggregator can charge from the consumer. If so, the total profit of the aggregator equals

$$\Pi^R(s) = \frac{1}{2}(p_1^{fb} + p_2^{fb})\bar{s} - P(x_1 + s)s - P(x_2 + \bar{s} - s)(\bar{s} - s) + R(s)$$

if it withdraws s from the grid in the first period. The final term in the above profit expression is a compensation payment defined in (8) that is levied on consumers in a lump-sum fashion. The marginal incentive to increase period 1 electricity consumption equals

$$\Pi^{R'}(s) = P(x_2 + \bar{s} - s) - P(x_1 + s)$$

by the construction of $R(s)$. The marginal payment $R'(s)$ neutralizes the aggregator's monopoly power in the real-time market. Moreover, the compensation payment is designed in such a way that the actual payment is zero (balanced-budget) at the first-best efficient allocation, i.e. $R(s^{fb}) = 0$. The aggregator's profit in equilibrium equals $\Pi^R(s^{fb}) = (p_1^{fb} - p_2^{fb})(\frac{1}{2}\bar{s} - s^{fb})$, which is non-negative; see Corollary 1. I state the following immediate result without proof:

Proposition 7 *The compensation function*

$$R(s) = \int_{s^{fb}}^s [P'(x_1 + y)y - P'(x_2 + \bar{s} - y)(\bar{s} - y)]dy \stackrel{\geq}{\leq} 0 \quad (8)$$

incites a structurally independent aggregator with a monopoly in the market for flexible distribution resources to withdraw the first-best efficient amount s^{fb} of electricity from the grid.

The compensation payment is zero in equilibrium, $R(s^{fb}) = 0$, and satisfies the aggregator's participation constraint, $\Pi^R(s^{fb}) \geq 0$.

Figure 2 essentially reproduces Figure 1 to illustrate how the compensation payment $R(s)$ operates in this framework. As in Figure 1, period 2 is the peak demand period by construction, $x_2 > x_1$, so that an unregulated aggregator in equilibrium withdraws relatively more electricity from the grid in period 1 than in period 2, $s^A > \frac{1}{2}\bar{s}$. Yet, exploitation of market power implies that electricity consumption is downward distorted in the off-peak period, $s^A < s^{fb}$.

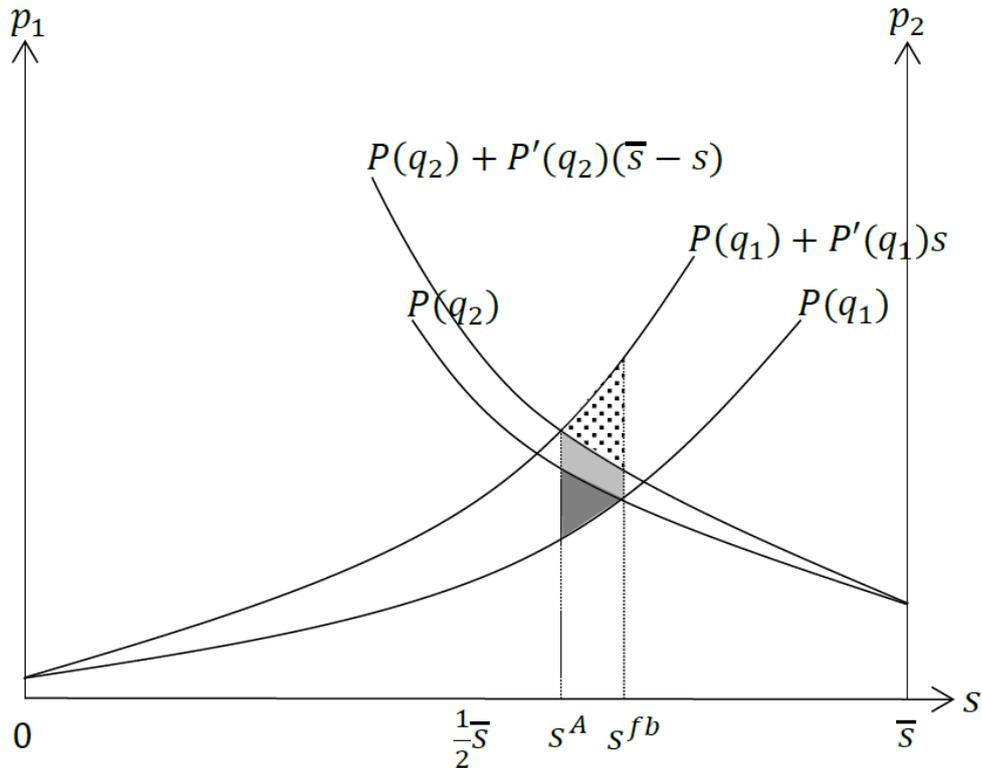


Figure 2: Regulation

Increasing electricity withdrawal in the first period from s^A to the first-best efficient level s^{fb} increases the aggregator's total expenditures in the real-time market by the dotted area in Figure 2 and thus is unprofitable without any compensation. An increase in electricity consumption in period 1 from s^A to s^{fb} under $R(s)$ increases the first period compensation by an area equal to the dotted plus the medium grey and the dark grey area in Figure 2. Period 2 compensation falls by the medium grey area. By summing up all incremental effects, it follows that an increase from s^A to s^{fb} increases the aggregator's profit by the dark grey area in the figure. This is equal to the total efficiency gain of increasing electricity consumption in the first period up to the efficient level from the profit maximizing level. Turning this argument around, the monopoly aggregator is residual party to all efficiency losses associated with a deviation from s^{fb} by the construction of $R(s)$.

Implementation of $R(s)$ requires that the supply functions of all participants in the real-time market are observable by the regulatory authority. The regulator can often collect the necessary data to compute these functions from the DSO operating the real-time market. Implementing $R(s)$ can be more complicated in other settings. A relevant example is when the DSO itself participates in the market for flexible distribution resources. In this case, the DSO probably has weak incentives to truthfully disclose the supply functions in the real-time market because doing so might enable the regulator to extract the DSO's full surplus. In this instance, the regulatory authority should collect bid data directly from the participants in the real-time market to increase regulatory efficiency.

4 Policy discussion

This paper has built a simple two-period model of an electricity market in which smart grid solutions enable third parties to supply flexible distribution resources (FDRs), in the form of intertemporal substitution of household electricity consumption, for the purpose of facilitating local short-term balancing of production and consumption. Examples of such consumption include indoor temperature control (heating or cooling) or battery charge and discharge. The main purpose was to examine how assumptions about the structure of the market for FDRs affect the efficiency with which these resources are deployed in the balancing market. Despite the simplicity of the model, it has generated results and policy implications that are likely to be robust and carry over to more general settings.

At the first-best efficient solution, FDRs are allocated in such a way as to equate real-time prices across periods if the real-time market is otherwise competitive. If there is not enough substitution capacity to achieve price equalization, then efficiency dictates that all consumption take place in the off-peak demand period. However, control over substantial FDR capacity enables whoever controls this resource to wield market power. In particular, exploitation of market power causes firms to consume too much electricity in the peak relative to the off-peak demand period. This means that resources are sometimes used in such a way as to create peak and off-peak price differences unrelated to any real scarcity problems in the market.

The market power problem is smaller if a larger number of firms, *aggregators*, compete for FDRs. The real-time market is perfectly competitive in the limit as the number of such aggregators becomes very large. This market performance depends on structural independence in the sense that aggregators' only role is to participate in the market for FDRs. I show in an example how an aggregator with monopoly power that also supplies thermal electricity to the real-time market, will deploy resources less efficiently than in the benchmark case without *any* FDR market. Such integration may be unavoidable with free entry into the FDR market, in which case price regulation of FDR supply may be called for. I characterize a regulatory policy that can implement the first-best efficient allocation on the basis of data that can be obtained from the day-ahead and real-time market.

Reaping the competitive benefits of an FDR market with multiple aggregators requires a formal local real-time market, known as a flexi market, where aggregators can bid in their

capacity. A key policy question is whether the distribution system operator (DSO) responsible for operating this market should also be allowed to supply FDR resources. In this case, where system operation is intertwined with the DSO's operations in the FDR market, the DSO has an incentive to distort allocations in the real-time market either directly or through the information it releases about that market. The resulting outcome is less efficient than a solution without any FDR market. Efficiency concerns in the real-time market then suggest that the DSO should not be allowed to participate in the FDR market, even if direct foreclosure of third parties is not a problem.

A market design featuring formal real-time markets at the control area level is not necessarily economically viable, for instance if running a local real-time market is associated with scale returns that may not be achieved in a small control area. Absent a formalized real-time market, perhaps the only option is to delegate the supply of flexible distribution resources to the DSO itself. By implication, the DSO becomes a monopolist in the FDR market. This monopoly power cannot be mitigated by standard revenue regulation because the exercise of market power stems from an incentive to minimize expenditures to balance the market. Yet, DSO monopoly power is not an argument against developing a market for flexible distribution resources because efficiency is still higher than in the benchmark case where those resources go unused.

The model has been cast in a framework in which firms exploit market power regarding flexible distribution resources, but all other supply is bid into the market at marginal cost. Allowing also other firms to exercise market power would probably not alter the welfare ranking of the different market structures. For instance, market performance would in all likelihood be higher under a set of structurally independent aggregators compared to the case of a monopoly firm operating in the market for FDRs. In fact, imperfect competition in the supply of balancing power would most likely amplify some of the welfare comparisons. The reason is that more efficient deployment of FDRs would increase the price elasticity of demand and therefore improve competition also on the supply side.

References

- Brown, David P. and David E.M. Sappington (2018): Optimal Procurement of Distributed Energy Resources. *The Energy Journal* 39(5), 131–155.
- Brown, David P. and David E.M. Sappington (2019): Employing Cost Sharing to Motivate the Efficient Implementation of Distributed Energy Resources. *Energy Economics* 81(June), 974–1001.
- Burger, Scott, Jose Pablo Chaves-Ávila, Carlos Batlle and Ignacio J. Pérez-Arriaga (2017): A Review of the Value of Aggregators in Electricity Systems. *Renewable and Sustainable Energy Reviews* 77(September), 395–405.
- Burger, Scott P., Jesse D. Jenkins, Carlos Batlle and Ignacio J. Pérez-Arriaga (2019a): Restructuring Revisited Part 1: Competition in Electricity Distribution Systems. *The Energy Journal* 40(3), 31–54.
- Burger, Scott P., Jesse D. Jenkins, Carlos Batlle and Ignacio J. Pérez-Arriaga (2019b): Restructuring Revisited Part 2: Coordination in Electricity Distribution Systems. *The Energy Journal* 40(3), 55–76.
- Bushnell, James B., Erin T. Mansur and Celeste Saravia (2008): Vertical Arrangements, Market Structure, and Competition: An Analysis of Restructured US Electricity Markets. *American Economic Review* 98(1), 237–66.
- EDSO (2015): Coordination of Transmission and Distribution System Operators: A Key Step for the Energy Union, <https://www.edsoforsmartgrids.eu/wp-content/uploads/public/Coordination-of-transmission-and-distribution-system-operators-May-2015.pdf>, October 15, 2019.
- Kim, Seung Wan, Michael G. Pollitt, Young Gyu Jin, Jip Kim and Yong Tae Yoon (2017): Contractual Framework for the Devolution of System Balancing Responsibility from the Transmission System Operator to Distribution System Operators. University of Cambridge, EPRG Working Paper 1715.
- Pollitt, Michael G. and Karim L. Anaya (2019): Competition in Markets for Ancillary Services? The Implications of Rising Distributed Generation. University of Cambridge, EPRG Working Paper 1928.
- Rey, Patrick and Jean Tirole (2007): A Primer on Foreclosure. In Mark Armstrong and Robert Porter (eds): *Handbook of Industrial Organization Volume 3*. Amsterdam, The Netherlands: North-Holland, 2145–2220.
- Tangerås, Thomas and Johannes Mauritzen (2018): Real-time versus Day-ahead Market Power in a Hydro-based Electricity Market. *Journal of Industrial Economics* 66(4), 904–941.
- Wolak, Frank A. (2007): Quantifying the Supply-side Benefits from Forward Contracting in Wholesale Electricity Markets. *Journal of Applied Econometrics* 22(7), 1179–1209.

Appendix

Proof of Lemma 1

The Lagrangian $-C(q_1) - C(q_2) + \underline{\lambda}s + \bar{\lambda}(\bar{s} - s)$ of the central planner's problem is strictly concave by the properties of $C(q)$, where $\underline{\lambda}$ is the Kuhn-Tucker (KT) multiplier associated with $s \geq 0$, and $\bar{\lambda}$ is the KT multiplier associated with $s \leq \bar{s}$. The first-order condition

$$-C'(q_1^{fb}) + C'(q_2^{fb}) + \underline{\lambda}^{fb} - \bar{\lambda}^{fb} = 0 \quad (9)$$

and complementary slackness conditions

$$s^{fb} \in [0, \bar{s}], \underline{\lambda}^{fb} \geq 0, \bar{\lambda}^{fb} \geq 0, \underline{\lambda}^{fb}s^{fb} = \bar{\lambda}^{fb}(\bar{s} - s^{fb}) = 0 \quad (10)$$

are therefore necessary and sufficient to characterize the unique first-best efficient allocation $(s^{fb}, q_1^{fb}, q_2^{fb})$ and first-best efficient KT multipliers $(\underline{\lambda}^{fb}, \bar{\lambda}^{fb})$. It is straightforward to verify that the following are solutions to (9) and (10): If $|x_1 - x_2| \leq \bar{s}$, then $(s^{fb}, q_1^{fb}, q_2^{fb}) = (s^*, q^*, q^*)$ and $\underline{\lambda}^{fb} = \bar{\lambda}^{fb} = 0$; if $x_1 - x_2 > \bar{s}$, then $(s^{fb}, q_1^{fb}, q_2^{fb}) = (0, x_1, x_2 + \bar{s})$, $\underline{\lambda}^{fb} = C'(x_1) - C'(x_2 + \bar{s}) > 0$ and $\bar{\lambda}^{fb} = 0$; if $x_2 - x_1 > \bar{s}$, then $(s^{fb}, q_1^{fb}, q_2^{fb}) = (\bar{s}, x_1 + \bar{s}, x_2)$, $\underline{\lambda}^{fb} = 0$ and $\bar{\lambda}^{fb} = C'(x_2) - C'(x_1 + \bar{s}) > 0$.

If $x_1 - x_2 > \bar{s} > 0$, then $q_1^{fb} - q_2^{fb} = x_1 - x_2 - \bar{s} > 0$ and therefore $(q_1^{fb} - q_2^{fb})(x_1 - x_2) > 0$. If $x_2 - x_1 > \bar{s} > 0$, then $q_1^{fb} - q_2^{fb} = x_1 - x_2 + \bar{s} < 0$ and again $(q_1^{fb} - q_2^{fb})(x_1 - x_2) > 0$. ■

Proof of Proposition 2

The Lagrangian $\Pi^{DSO}(s) + \underline{\lambda}s + \bar{\lambda}(\bar{s} - s)$ of the DSO is strictly concave by the assumption of competitive supply of thermal electricity in the short-term market and the properties of $C'(\cdot)$. Hence, the first-order condition

$$p_2^{DSO} - p_1^{DSO} - P'(q_1^{DSO})q_1^{DSO} + P'(q_2^{DSO})q_1^{DSO} + \underline{\lambda}^{DSO} - \bar{\lambda}^{DSO} = 0 \quad (11)$$

and complementary slackness conditions

$$s^{DSO} \in [0, \bar{s}], \underline{\lambda}^{DSO} \geq 0, \bar{\lambda}^{DSO} \geq 0, \underline{\lambda}^{DSO}s^{DSO} = \bar{\lambda}^{DSO}(\bar{s} - s^{DSO}) = 0 \quad (12)$$

are necessary and sufficient optimality conditions for s^{DSO} and the equilibrium KT multipliers $(\underline{\lambda}^{DSO}, \bar{\lambda}^{DSO})$, where $p_i^{DSO} = P(q_i^{DSO})$ is the equilibrium real-time price in period $i = 1, 2$.

Part (i) of the proposition: If $x_1 - x_2 > \bar{s}$, then $(s^{DSO} - s^{fb})(x_1 - x_2) = s^{DSO}(x_1 - x_2) \geq 0$ because $s^{fb} = 0$ by Lemma 1. By that same Lemma, $(s^{DSO} - s^{fb})(x_1 - x_2) = (\bar{s} - s^{DSO})(x_2 - x_1) \geq 0$ for $x_2 - x_1 > \bar{s}$ because then $s^{fb} = \bar{s}$. If $x_1 = x_2$, then $(s^{DSO} - s^{fb})(x_1 - x_2) \geq 0$ trivially holds.

The final case is where $|x_1 - x_2| \in (0, \bar{s}]$. By full price equalization for $s^{fb} = s^* \in [0, \bar{s}]$:

$$\begin{aligned}\Pi^{DSO}(s^{fb}) &= -P(x_1 + s^{fb}) - P'(x_1 + s^{fb})(x_1 + s^{fb}) + P(x_2 + \bar{s} - s^{fb}) + P'(x_2 + \bar{s} - s^{fb})(x_2 + \bar{s} - s^{fb}) \\ &= -P'(x_1 + s^{fb})(x_1 + s^{fb}) + P'(x_2 + \bar{s} - s^{fb})(x_2 + \bar{s} - s^{fb}) \\ &= P'(q^*)(x_1 - x_2).\end{aligned}$$

As $P' > 0$, it follows that $\Pi^{DSO}(s^{fb}) > 0$ for $x_2 + \bar{s} \geq x_1 > x_2$, in which case $s^{DSO} > s^{fb}$ by strict concavity of $\Pi^{DSO}(s)$ and because $s^{fb} < \bar{s}$ in this range of (x_1, x_2) . In an analogous manner, $x_1 + \bar{s} \geq x_2 > x_1$ implies $s^{DSO} < s^{fb}$. Hence, $(s^{DSO} - s^{fb})(x_1 - x_2) > 0$ for $|x_1 - x_2| \in (0, \bar{s}]$.

Part (ii) of the proposition: Rearrange the first-order condition (11) and multiply through by $(p_1^{DSO} - p_2^{DSO})$ to get

$$\begin{aligned}[2P'(q_1^{DSO})(\frac{1}{2}\bar{s} - s^{DSO}) + \underline{\lambda}^{DSO} - \bar{\lambda}^{DSO}](p_1^{DSO} - p_2^{DSO}) \\ = [P'(q_1^{DSO}) - P'(q_2^{DSO})][p_1^{DSO} - p_2^{DSO}](\bar{s} - s^{DSO}) + (p_1^{DSO} - p_2^{DSO})^2.\end{aligned}\quad (13)$$

The first term on the right-hand side (RHS) of (13) is non-negative because convexity of the marginal production cost $C'(q)$ implies

$$[P'(x_1 + s) - P'(x_2 + \bar{s} - s)][P(x_1 + s) - P(x_2 + \bar{s} - s)] \geq 0 \text{ for all } s \in [0, \bar{s}], \quad (14)$$

The second term on the right-hand side (RHS) of (13) is strictly positive for all $p_1^{DSO} \neq p_2^{DSO}$. The left-hand side (LHS) of (13) is zero if $s^{DSO} = \frac{1}{2}\bar{s}$ by $\underline{\lambda}^{DSO} s^{DSO} = \bar{\lambda}^{DSO}(\bar{s} - s^{DSO}) = 0$. Hence, $p_1^{DSO} \neq p_2^{DSO}$ implies $s^{DSO} \neq \frac{1}{2}\bar{s}$. If $p_1^{DSO} > p_2^{DSO}$, then LHS of (13) is strictly negative if $s^{DSO} > \frac{1}{2}\bar{s}$ by $P'(q_1^{DSO}) > 0$, $\bar{\lambda}^{DSO} \geq 0$ and $\underline{\lambda}^{DSO} s^{DSO} = 0$. Hence, $p_1^{DSO} > p_2^{DSO}$ implies $s^{DSO} < \frac{1}{2}\bar{s}$. Similarly, $p_2^{DSO} > p_1^{DSO}$ implies $s^{DSO} > \frac{1}{2}\bar{s}$. Next, let $x_1 > x_2$ and suppose $s^{DSO} \geq \frac{1}{2}\bar{s}$. In this case,

$$q_1^{DSO} - q_2^{DSO} = x_1 - x_2 + 2s^{DSO} - \bar{s} > 0,$$

and therefore $p_1^{DSO} = P(q_1^{DSO}) > P(q_2^{DSO}) = p_2^{DSO}$ by $P' > 0$. But then $s^{DSO} < \frac{1}{2}\bar{s}$ from the previous argument, which is a contradiction. Hence, $x_1 > x_2$ implies $s^{DSO} < \frac{1}{2}\bar{s}$. By a similar argument, $x_2 > x_1$ implies $s^{DSO} > \frac{1}{2}\bar{s}$. This concludes the proof that $(\frac{1}{2}\bar{s} - s^{DSO})(x_1 - x_2) > 0$ for $x_1 \neq x_2$. ■

Proof of Corollary 2

Rearrange the first-order condition (11) as

$$2P'(q_1^{DSO})(\frac{1}{2}\bar{s} - s^{DSO}) + \underline{\lambda}^{DSO} - \bar{\lambda}^{DSO} = p_1^{DSO} - p_2^{DSO} + [P'(q_1^{DSO}) - P'(q_2^{DSO})](\bar{s} - s^{DSO}). \quad (15)$$

Assume that $x_1 > x_2$. I know from Proposition 2 that $s^{DSO} < \frac{1}{2}\bar{s}$ in this case. By implication, LHS of (15) is strictly positive since $P'(q_1^{DSO}) > 0$, $\underline{\lambda}^{DSO} \geq 0$ and $\bar{\lambda}^{DSO}(\bar{s} - s^{DSO}) = 0$. RHS of (15) is negative if $q_1^{DSO} < q_2^{DSO}$ because then $p_1^{DSO} = P(q_1^{DSO}) < P(q_2^{DSO}) = p_2^{DSO}$ by

$P' > 0$ and $P'(q_1^{DSO}) \leq P'(q_2^{DSO})$ by (14). Hence, $x_1 > x_2$ implies $q_1^{DSO} \geq q_2^{DSO}$. RHS of (15) is zero if $q_1^{DSO} = q_2^{DSO}$. I conclude that $x_1 > x_2$ implies $q_1^{DSO} > q_2^{DSO}$, and therefore $p_1^{DSO} = P(q_1^{DSO}) > P(q_2^{DSO}) = p_2^{DSO}$ by $P' > 0$. By analogous arguments, $x_2 > x_1$ implies $p_2^{DSO} > p_1^{DSO}$. ■

Proof of Proposition 3

The Lagrangian $\Pi^A(s_a) + \underline{\lambda}_a s_a + \bar{\lambda}_a(L_a \bar{s} - s_a)$ of aggregator a is strictly concave by the assumptions that thermal capacity is competitively supplied and the properties of $C'(\cdot)$. Hence, the first order-condition

$$p_2^A - p_1^A - \frac{1}{A}P'(q_1^A)s^A + \frac{1}{A}P'(q_2^A)(\bar{s} - s^A) + \underline{\lambda}^A - \bar{\lambda}^A = 0 \quad (16)$$

and complementary slackness conditions

$$s^A \in [0, \bar{s}], \underline{\lambda}^A \geq 0, \bar{\lambda}^A \geq 0, \underline{\lambda}^A s^A = \bar{\lambda}^A(\bar{s} - s^A) = 0 \quad (17)$$

characterize the unique symmetric second-stage equilibrium $s_a^A = s^A/A$, $L_a^A = 1/A$, $\bar{\lambda}_a^A = \bar{\lambda}^A$, $\underline{\lambda}_a^A = \underline{\lambda}^A$ for all a , and where $q_1^A = x_1 + s^A$, $q_2^A = x_1 + \bar{s} - s^A$ and $p_i^A = P(q_i^A)$, $i = 1, 2$.

Part (i) of the proposition: I omit the proof that $(s^{fb} - s^A)(x_1 - x_2) \geq 0$ with strict inequality if $|x_1 - x_2| \in (0, \bar{s}]$ because it is identical to the proof of the first part of Proposition 2. As for the comparative statics result, differentiate the equilibrium condition $\Pi^{A'}(\frac{1}{A}s^A) = 0$ for $s^A \in (0, \bar{s})$:

$$\frac{ds^A}{dA}(x_1 - x_2) = \frac{-(p_1^A - p_2^A)(x_1 - x_2)}{(A+1)(P'(q_1^A) + P'(q_2^A)) + P''(q_1^A)s^A + P''(q_2^A)(\bar{s} - s^A)}.$$

The denominator is strictly positive. I can then follow the same steps as in the proof of Corollary 2 to establish $(p_1^A - p_2^A)(x_1 - x_2) > 0$ for all $x_1 \neq x_2$.

Part (ii) of the proposition: Consider first necessity. Obviously, $s^A = s^{DSO}$, $\underline{\lambda}^A = \underline{\lambda}^{DSO}$ and $\bar{\lambda}^A = \bar{\lambda}^{DSO}$ satisfy (5) and (17) for $A = 1$. The necessity of $x_1 \neq x_2$ is trivial. For $s^{DSO} = \bar{s}$, it is straightforward to verify that $s^A = \bar{s}$, $\underline{\lambda}^A = 0$ and

$$\bar{\lambda}^A = \frac{A-1}{A}(P(x_2) - P(x_1 + \bar{s})) + \frac{1}{A}\bar{\lambda}^{DSO} \geq 0$$

solve (5) and (17). To see why $\bar{\lambda}^A \geq 0$ in this case, recall that $(p_1^{DSO} - p_2^{DSO})(\frac{1}{2}\bar{s} - s^{DSO}) \geq 0$ from the proof of Proposition 2. Hence, $s^{DSO} = \bar{s}$ implies $p_2^{DSO} - p_1^{DSO} = P(x_2) - P(x_1 + \bar{s}) \geq 0$. Similarly, $s^{DSO} = 0$ implies that $s^A = 0$, $\bar{\lambda}^A = 0$ and

$$\underline{\lambda}^A = \frac{A-1}{A}(P(x_1) - P(x_2 + \bar{s})) + \frac{1}{A}\underline{\lambda}^{DSO} \geq 0$$

solve (5) and (17).

Consider next sufficiency, and assume that $A > 1$, $x_1 \neq x_2$ and $s^{DSO} \in (0, \bar{s})$. Evaluated at $s_{a'} = \frac{1}{A}s^{DSO}$ for all a' , the marginal profit of aggregator a simplifies to $\Pi^{A'}(\frac{1}{A}s^{DSO})(x_1 - x_2) = -\frac{A-1}{A}(p_1^{DSO} - p_2^{DSO})(x_1 - x_2) < 0$. The strict negativity follows from $(p_1^{DSO} - p_2^{DSO})(x_1 - x_2) > 0$

for all $x_1 \neq x_2$, see the proof of Corollary 2. It follows that $s^A < s^{DSO}$ if $x_1 > x_2$, whereas $s^A > s^{DSO}$ if $x_1 < x_2$. ■

Proof of Proposition 4

The problem of maximizing the total profit function $\Pi(\tilde{s}^{DSO}, \tilde{s}^A, x_1, x_2)$ characterized in (6) over $(\tilde{s}^{DSO}, \tilde{s}^A) \in [0, \bar{s}^{DSO}] \times [0, \bar{s}^A]$ is complicated by the fact that it is non-concave. Although $\Pi(s^{DSO}, s^A, x_1, x_2)$ is strictly concave in each of its separate arguments s^{DSO} and s^A , the Hessian matrix of $\Pi(s^{DSO}, s^A, x_1, x_2)$ has one negative and one positive Eigenvalue. This means that all local interior solutions (s^{DSO}, s^A) are saddle points. By implication, the optimal solution features corner solutions. Suppose $s^A \in (0, \bar{s}^A)$, in which case the necessary first-order condition $\frac{\partial \Pi}{\partial s^A} = 0$ implies

$$P'(q_2)(\bar{s}^{DSO} - s^{DSO}) - P'(q_1)s^{DSO} = 0. \quad (18)$$

By the assumption that $s^A \in (0, \bar{s}^A)$, either $s^{DSO} = 0$, in which case the LHS of (13) is strictly positive, or $s^{DSO} = \bar{s}^{DSO}$, in which case the RHS of (18) is strictly negative, both of which violate condition (18). Thus, the equilibrium features $s^A \in \{0, \bar{s}^A\}$. Let $\pi(s^A, x_1, x_2) = \max_{\tilde{s}^{DSO} \in [0, \bar{s}^{DSO}]} \Pi(\tilde{s}^{DSO}, s^A, x_1, x_2)$. By definition, $s^A = \bar{s}^A$ if $\pi(\bar{s}^A, x_1, x_2) > \pi(0, x_1, x_2)$, $s^A = 0$ if the strict inequality is reversed, and the DSO is indifferent between $s^A = \bar{s}^A$ and $s^A = 0$ if $\pi(\bar{s}^A, x_1, x_2) = \pi(0, x_1, x_2)$.

Claim 3 If $\frac{\partial \Pi(\bar{s}^{DSO}, 0, x_1, x_2)}{\partial s^{DSO}} \geq 0$ for all $x_1 \geq x_2$, then $s^A = \bar{s}^A$ for all $x_1 > x_2$.

Proof: By this assumption, $s^{DSO}(0, x_1, x_2) = \bar{s}^{DSO}$ and therefore $\pi(0, x_1, x_2) = -P(x_1 + \bar{s}^{DSO})\bar{s}^{DSO}$ for all $x_1 \geq x_2$. Then

$$\pi(\bar{s}^A, x_1, x_2) - \pi(0, x_1, x_2) \geq \Pi(0, \bar{s}^A, x_1, x_2) - \pi(0, x_1, x_2) = [P(x_1 + \bar{s}^{DSO}) - P(x_2 + \bar{s}^{DSO})]\bar{s}^{DSO} > 0$$

for all $x_1 > x_2$ by $P' > 0$. ■

Claim 4 If $\frac{\partial \Pi(\bar{s}^{DSO}, 0, x_2, x_2)}{\partial s^{DSO}} \leq 0$, then $s^A = \bar{s}^A$ for all $x_1 > x_2$.

Proof: Seeing as $\frac{\partial \Pi(s^{DSO}, s^A, x_1, x_2)}{\partial s^{DSO} \partial s^A} < 0$ and $\frac{\partial \Pi(s^{DSO}, s^A, x_1, x_2)}{\partial s^{DSO} \partial x_1} < 0$ for all s^{DSO} and s^A , it follows that $s^{DSO}(s^A, x_1, x_2) < \bar{s}^{DSO}$ for all $s^A \in [0, \bar{s}^A]$ and $x_1 > x_2$ under the assumed properties of this claim. Moreover,

$$\frac{\partial \Pi(0, 0, x_2, x_2)}{\partial s^{DSO}} = P(x_2 + \bar{s}) - P(x_2) + P'(x_2 + \bar{s})\bar{s}^{DSO} > 0$$

implies $s^{DSO}(0, x_2, x_2) > 0$. By continuity, $s^{DSO}(s^A, x_1, x_2) > 0$ also for a subset $s^A > 0$ and $x_1 > x_2$ with positive measure. Next,

$$\frac{\partial \pi(s^A, x_1, x_2)}{\partial s^A} = P'(q_2)(\bar{s}^{DSO} - s^{DSO}) - P'(q_1)s^{DSO}$$

by the envelope theorem, and

$$\frac{\partial^2 \pi(s^A, x_1, x_2)}{\partial s^A \partial x_1} = -[P'(q_1) + P'(q_2) + P''(q_2)(\bar{s}^{DSO} - s^{DSO})] \frac{\partial s^{DSO}}{\partial x_1} - P''(q_1) s^{DSO} \left(1 + \frac{\partial s^{DSO}}{\partial x_1}\right),$$

where

$$\frac{\partial s^{DSO}}{\partial x_1} = \frac{-[P'(q_1) + P''(q_1) s^{DSO}]}{2P'(q_1) + 2P'(q_2) + P''(q_1) s^{DSO} + P''(q_2)(\bar{s}^{DSO} - s^{DSO})} < 0$$

for all $s^{DSO} \in (0, \bar{s}^{DSO})$. Observe that $\frac{\partial^2 \pi(s^A, x_1, x_2)}{\partial s^A \partial x_1} = 0$ for $s^{DSO} = 0$ and

$$\frac{\partial^2 \pi(s^A, x_1, x_2)}{\partial s^A \partial x_1} = P'(q_1) \frac{P'(q_1) + P'(q_2) \frac{s^{DSO}}{q_1} \left[\frac{x_1 + s^A}{\bar{s}^{DSO}} + 1 - \frac{P''(q_1) q_1}{P'(q_1)}\right] + P''(q_2)(\bar{s}^{DSO} - s^{DSO})}{2P'(q_1) + 2P'(q_2) + P''(q_1) s^{DSO} + P''(q_2)(\bar{s}^{DSO} - s^{DSO})} > 0$$

for $s^{DSO} \in (0, \bar{s}^{DSO})$ yield

$$\pi(\bar{s}^A, x_1, x_2) - \pi(0, x_1, x_2) - [\pi(\bar{s}^A, x_2, x_2) - \pi(0, x_2, x_2)] = \int_0^{\bar{s}^A} \int_{x_2}^{x_1} \frac{\partial^2 \pi(s^A, y, x_2)}{\partial s^A \partial x_1} dy ds^A > 0$$

for all $x_1 > x_2$. Finally,

$$\pi(\bar{s}^A, x_2, x_2) - \pi(0, x_2, x_2) \geq \Pi(\bar{s}^{DSO} - s^{DSO}(0, x_2, x_2), \bar{s}^A, x_2, x_2) - \pi(0, x_2, x_2) = 0$$

completes the proof. ■

Claim 5 If $\frac{\partial \Pi(\bar{s}^{DSO}, 0, x_2, x_2)}{\partial s^{DSO}} > 0$ and $\frac{\partial \Pi(\bar{s}^{DSO}, 0, x_1^c, x_2)}{\partial s^{DSO}} = 0$ for some $x_1^c > x_2$, then $s^A = \bar{s}^A$ for all $x_1 > x_2$.

Proof: By a line of argument similar to the one used to prove Claim 1, it follows that $\pi(\bar{s}^A, x_1, x_2) > \pi(0, x_1, x_2)$ for all $x_1 \in (x_2, x_1^c]$. If $x_1 > x_1^c$, then

$$\pi(\bar{s}^A, x_1, x_2) - \pi(0, x_1, x_2) - [\pi(\bar{s}^A, x_1^c, x_2) - \pi(0, x_1^c, x_2)] = \int_0^{\bar{s}^A} \int_{x_1^c}^{x_1} \frac{\partial^2 \pi(s^A, y, x_2)}{\partial s^A \partial x_1} dy ds^A \geq 0$$

because $s^{DSO}(s^A, x_1, x_2) < \bar{s}^{DSO}$ for all $s^A \in [0, \bar{s}^A]$ and $x_1 > x_1^c$. Combining this inequality with $\pi(\bar{s}^A, x_1^c, x_2) > \pi(0, x_1^c, x_2)$ concludes the proof of the claim. ■

Summarizing the above three claims yields $s^A = \bar{s}^A$ for all $x_1 > x_2$. By following qualitatively similar steps as the above, it is straightforward to verify that $s^A = 0$ for all $x_1 < x_2$. If $x_1 > x_2$, then $s^{DSO} + s^A - \frac{1}{2}\bar{s} = s^{DSO} + \frac{1}{2}(\bar{s}^A - \bar{s}^{DSO})$, which is non-negative if $\bar{s}^{DSO} \leq \bar{s}^A$. If $x_1 < x_2$, then $s^{DSO} + s^A - \frac{1}{2}\bar{s} = s^{DSO} - \bar{s}^{DSO} - \frac{1}{2}(\bar{s}^A - \bar{s}^{DSO})$, which is non-positive if $\bar{s}^{DSO} \leq \bar{s}^A$. Hence, $\bar{s}^{DSO} \leq \bar{s}^A$ implies $(s^{DSO} + s^A - \frac{1}{2}\bar{s})(x_1 - x_2) \geq 0$ for all (x_1, x_2) . ■

Proof of Proposition 5

The Lagrangian $\Pi^I(s) + \underline{\lambda}s + \bar{\lambda}(\bar{s} - s)$ of the integrated firm is strictly concave by the properties of $C'(\cdot)$. Hence, the profit-maximizing solution $(s^I, \underline{\lambda}^I, \bar{\lambda}^I)$ is uniquely determined by the first-order

condition

$$p_2^I - p_1^I + P'(q_1^I)x_1 - P'(q_2^I)x_2 + \underline{\lambda}^I - \bar{\lambda}^I = 0 \quad (19)$$

and complementary slackness conditions

$$s^I \in [0, \bar{s}], \underline{\lambda}^I \geq 0, \bar{\lambda}^I \geq 0, \underline{\lambda}^I s^I = \bar{\lambda}^I (\bar{s} - s^I) = 0, \quad (20)$$

where $q_1^I = x_1 + s^I$, $q_2^I = x_2 + \bar{s} - s^I$ and $p_i^I = P(q_i^I)$, $i = 1, 2$. Observe that $(s^I, \underline{\lambda}^I, \bar{\lambda}^I) = (\frac{1}{2}\bar{s}, 0, 0)$ solves (19) and (20) for $x_1 = x_2$. Hence,

$$s^I(x_1, x_2) - \frac{1}{2}\bar{s} = \int_{x_2}^{x_1} \frac{\partial s^I(y, x_2)}{\partial x_1} dy$$

$(s^I(x_1, x_2) - \frac{1}{2}\bar{s})(x_1 - x_2) \geq 0$ with strict inequality if $x_1 \neq x_2$ and $C'''(\cdot) > 0$ then follows from

$$\frac{\partial s^I(y, x_2)}{\partial x_1} = \frac{C'''(x_1 + s^I(y, x_2))y}{-\Pi^{I''}(s^I(y, x_2))} \text{ for all } s^I(y, x_2) \in (0, \bar{s}).$$

■