An Optimal Pricing Theory of Transaction Fees in Thin Markets *

Simon Loertscher† Andras Niedermayer‡

January 5, 2019

Note: Online Appendices begin on page 43.

Abstract

Transaction fees and taxes fare prominently in policy debates. We derive optimal transaction fees as the solution to a pricing problem faced by intermediaries and show that no mechanism performs better. We show that a higher elasticity of supply and a larger weight on the intermediary's profits lead to higher fees. Counterintuitively, more elastic demand may increase fees. We show that extreme value theory implies asymptotic optimality of linear fees in thin markets. Our theoretical results fit empirical observations in several industries with intermediaries.

Keywords: brokerage, fee-setting, percentage fees, thin markets, Pareto distributions.

JEL-Classification: C72, C78, L13

*We want to thank Arnaud Costinot, Julio González-Díaz, David Genesove, Philipp Kircher, Stephan Lauermann, Volker Nocke, François Ortalo-Magné, Marco Ottaviani, Martin Peitz, Bernard Salanié, Juan Santaló, Mark Satterthwaite, Philipp Schmidt-Dengler, Art Shneyerov, Dan Spulber, Adam Szeidl, Ernst-Ludwig von Thadden, Rakesh Vohra, Cédric Wasser, Asher Wolinsky, Abdullah Yavas, participants at the IIOC 2012 in Washington D.C., the SFB TR15 2012 meeting in Mannheim, the Swiss IO Day 2012 in Bern, EARIE 2012 in Rome, and seminar audiences at Bates White, Bristol, Columbia, Deakin, Georgetown, Hamburg, LSE, Mannheim, Melbourne, Michigan, Munich, UNSW, Northwestern, Sydney, WHU, Zurich, and the University of Pennsylvania for their comments and suggestions. We want to thank Mark Satterthwaite for stating the conjecture of convergence to linear fees as the opportunity costs of trading increase and for having encouraged us to prove this result. Financial support through a research grant by the Faculty of Business and Economics at the University of Melbourne is also gratefully acknowledged. The second author gratefully acknowledges financial support by the German Science Foundation through SFB TR-15 and project PE 813/2-2.

†Department of Economics, Level 4, FBE Building, University of Melbourne, 111 Barry St, Victoria 3010, Australia. Email: simonl@unimelb.edu.au.

‡Université Paris-Dauphine, PSL Research University, LEDa, 75016 Paris, France. Email: andras.niedermayer@dauphine.fr.

1 Introduction

Internet platforms, brokers, and governments around the globe charge commission fees and taxes that are paid when a transaction occurs. Examples include the indirect taxes imposed by governments, the percentage fees charged by real-estate brokers, headhunters, and stock brokers, and the commission fees charged by auction houses and trading platforms such as Sotheby's, Christie's, eBay, iTunes, and Amazon. In the United States alone, the annual volume of trade via intermediaries who charge fees is in the order of trillions of USD.[1] Transaction fees also fare prominently in public policy debates on issues as various as credit-card fees, allegations of collusive commission fee-setting by auction houses and real-estate agents, antitrust investigations into pricing by internet platforms, the – at times, drastic – increases of value-added taxes in the wake of the global financial crisis, or the introduction of financial transaction taxes in the European Union (EU).[2] Little is known to date about the effects and determinants of the structure of such fees and as to when their use dominates alternative trading mechanisms.

In this paper, we analyze the fee structure from an optimal pricing perspective. Our focus is on thin markets in which every seller owns a unique object and faces only a small number of buyers in every period. Assuming independent private values on both the seller's and the buyers' sides, we show that there is no loss of generality by focusing on fee-setting insofar as no mechanism fares better. Interestingly, more elastic demand may lead to higher optimal fees, and linear fees are asymptotically optimal in increasingly thin markets.

We show a connection between the asymptotic optimality of linear fees in thin markets and asymptotic results from extreme value theory. This connection is deep and

[1] See, for example, Department of Justice (2007) for real-estate brokers and Business Wire, February 2014: "U.S. 2013 Credit and Debit Card Purchases Increased 8% Over 2012: Break $4 Trillion Barrier for First Time" for credit and debit cards.

[2] The European Commission published its revised proposal to introduce a financial transaction tax on February 14, 2013 (see *The Economist* on February 23, 2013: "Bin it - Europe's financial-transactions tax"). A previous version of the proposal already had been approved by the European Parliament and the Council of the European Union. Notwithstanding the Tobin tax as the standard economics rationale for such a measure, the proposal for a financial-transactions tax has been primarily motivated by a desire or need to increase tax revenue. For Amazon, which uses linear fees in 33 of the 38 categories it uses for third-party sellers, revenue from the long tail of the distributions appears to be quantitatively important.

novel. As markets become increasingly thin, the distributions of participants' valuations converge to Generalized Pareto distributions. This in turn implies convergence to linear fees. Besides providing an explanation for the linear fees often observed in practice, our asymptotic results allow for a surprisingly tractable analysis for an otherwise intractable problem.

While the result that more elastic demand can lead to higher optimal fees is counterintuitive, this and other results can be explained with concepts from monopoly and monopsony pricing. The reason for the counterintuitive demand-elasiticity effect is that the seller's price responds endogenously to changes in demand. When this reaction is excessive, the optimal fee needs to increase to partly offset this price endogeneity effect. For the limiting linear fees, the seller's and the intermediary's pricing incentives are perfectly aligned, so that the optimal fee is independent of the elasticity of demand.

Our paper contributes, first and foremost, to the growing literature that applies insights and methods from mechanism design to pertinent questions in industrial organization. Recent and complementary contributions, such as Board (2008), Gomes (2014), Tirole (2016) and Garrett (2016), have applied multi-period mechanism design to intertemporal pricing and to optimal incentive schemes for platform participation. Our paper, and the predecessor (Loertscher and Niedermayer, 2007) it builds on and supersedes, is the first paper to connect fee-setting to optimal pricing in thin markets with two-sided private information as first studied by Myerson and Satterthwaite (1983). In the companion paper, Loertscher and Niedermayer (2017), we take our model to the data.[3]

This combination of market thinness and two-sided private information is also what sets our theory apart from the existing theoretical literature on the transaction fees of profit maximizing intermediaries (Yavas (1992), Caillaud and Jullien (2003), Hagiu (2007), Matros and Zapechelnyuk (2008), Shy and Wang (2011), Niedermayer and Shneyerov (2014), Johnson (2014), and Wang and Wright (2017)) and on the indirect taxes charged by governments (Salanié (2003, Chapter 3), Delipalla and Keen (1992), and Anderson

[3] In Loertscher and Niedermayer (2018), we analyze transaction fees in a thick markets and show that they may be used in equilibrium to deter entry of a competing exchange.

et al. (2001a,b)).[4] Without this combination, the theory would be silent about the functional form of the fee, that is, as to whether it is fixed, a percentage fee or a non-linear fee. This highlights a robustness of linear fees: while in thin markets they are needed for optimality, in thick markets they do no harm, being equivalent to alternative ways of raising revenues. Moreover, our model predicts equilibrium price dispersion, which is consistent with the data but absent in most of the aforementioned models. We provide a more detailed discussion later on in the paper.

Optimal mechanisms for profit maximizing intermediaries in thin markets are, for example, studied by Myerson and Satterthwaite (1983), Baliga and Vohra (2003), and Jullien and Mariotti (2006). To the best of our knowledge, none of the existing papers contains an analysis of transaction fees, their structure, or their convergence to linearity.[5]

The remainder of this paper is organized as follows. Section 2 sets up the model, which is analyzed in Section 3. Section 4 provides the thin market analysis based on extreme value theory. Section 5 discusses microfoundations for transaction costs, comparative statics, empirical implications, and extensions. It also explains why the theory of optimal

[4] Yavas (1992), Caillaud and Jullien (2003), Shy and Wang (2011), Johnson (2014) and Wang and Wright (2017) assume that the seller's cost is public information (or, equivalently, that either there is no uncertainty about the seller's cost or that there is perfect competition between sellers). In Matros and Zapechelnyuk (2008) the seller's cost is sunk after he chooses to go to the intermediary, which can be best seen by considering the one period version of their model. Therefore, the seller's private cost only matters for his participation decision, but not for anything that happens after he chooses to participate (in particular for the reserve and transaction price). In Niedermayer and Shneyerov (2014) there is a continuum of sellers and buyers, so that by the law of large numbers there is no uncertainty about the realized distribution of sellers' costs. None of these models can account for the counterintuitive effect of the elasticity of demand on the equilibrium fees. Salanié (2003, Chapter 3) provides an overview of the literature on indirect taxes in competitive (that is, thick) markets. Delipalla and Keen (1992); Anderson et al. (2001a,b) consider optimal taxation with imperfect competition and *public* information about the seller's cost. The lack of relevant two-sided private information is what leads to the finding in these articles that optimal fees or taxes are higher if demand is *less* elastic. Moreover, models that assume thick markets generate an irrelevance result concerning the functional form of the fee or tax (fixed, percentage, linear, or non-linear), because absent any uncertainty about the seller's cost, the optimal mechanism for the intermediary is to set the seller indifferent and choose optimal one-sided pricing for buyers.

[5] Myerson and Satterthwaite (1983) are mostly cited for their impossibility result, but the paper actually also contains a section about a profit maximizing intermediary in a static bilateral trade market. Baliga and Vohra (2003) focus on market research rather than fees in a static model. Jullien and Mariotti (2006) assume two-sided private information in a static model with one broker and two buyers, but focus on fees that are a function of the reserve price rather than the transaction price without specifying the functional form of these fees. Moreover, as mentioned, we have multiple buyers and multiple periods. In the companion paper, we also structurally estimate the model.

linear pricing sheds little light on the optimality of linear fees. Section 6 concludes. Proofs and additional background material are in the Appendix.

2 Model

Motivated by the widespread use of fee-setting in intermediated markets with both long-term and spot contracts, we set up and analyze a general infinite horizon model. Time is discrete and indexed by $t = 0, 1, ...$ The basic analysis assumes that there is one intermediary and one seller. We provide extensions which relax this. The seller's *primitive* cost c_0 is the seller's private information and drawn from the primitive distribution G_0 with support $[\underline{c}_0, \overline{c}_0]$ and density $g_0(c_0) > 0$ for all $c_0 \in (\underline{c}_0, \overline{c}_0)$. His value of the outside option of not participating is zero. The cost c_0 can equivalently be thought of as the opportunity cost of selling or as a cost of production, both accruing to the seller in the period he sells.[6] The seller and the intermediary have the common discount factor $\delta \in [0, 1)$, which may represent time preferences or the period-to-period probability that the seller stays in the market as in Satterthwaite and Shneyerov (2008), or a combination thereof. Setting $\delta = 0$ yields the static model as a special case.

We assume that in every period there is a fixed number of potential buyers $\overline{B}$, each of whom enters with the independent probability $\tilde{\pi}$, so that the probability π_B of having exactly $B \leq \overline{B}$ buyers is given by the probability mass function for the binomial $\pi_B = \begin{pmatrix} \overline{B} \\ B \end{pmatrix} \tilde{\pi}^B (1 - \tilde{\pi})^{\overline{B}-B}$. Buyers who participate are sometimes also called bidders. Each bidder draws her primitive valuation v_0 independently from the (primitive) distribution F_0 with support $[\underline{v}_0, \overline{v}_0]$ and density $f_0(v_0) > 0$ for all $v_0 \in (\underline{v}_0, \overline{v}_0)$. The value of the outside option of not participating is zero for all buyers. All players – buyers, the seller, and the intermediary – are risk-neutral.

We call the buyer's valuation v_0 his *primitive valuation* as we assume that there are additional transaction costs, such as shipping costs, the option value of buying from another source in the future, and, specifically for real estate, moving costs and the

[6] For example, if the good is a real-estate property, the opportunity cost of selling is given by the discounted stream of income from renting the property or the discounted value of the flow utility from using the property.

opportunity of renting rather than buying. The *effective valuation* $v := K^B + \hat{K}^B v_0$ takes into account these transaction costs, where K^B should be thought of as type independent transaction costs (such as shipping costs) and $\hat{K}^B$ as type dependent costs (such as the opportunity cost of renting). With the exception of Section 4, we will treat the cost parameters K^B and $\hat{K}^B$ as exogenous and fixed and simplify notation by dealing with the effective valuation v and its corresponding distribution F with support $[\underline{v}, \overline{v}]$. Analogously, on the seller's side denote the seller's *effective costs* as $c := K^S + \hat{K}^S c_0$ with the corresponding distribution G and support $[\underline{c}, \overline{c}]$.[7]

Denoting by f and g the densities of F and G, respectively, we assume that the functions

$$\Phi(v) := v - \frac{1 - F(v)}{f(v)} \quad \text{and} \quad \Gamma(c) := c + \frac{G(c)}{g(c)}$$

are monotonically increasing in their arguments and continuously differentiable. Following Myerson (1981), $\Phi(v)$ is often called the virtual valuation while $\Gamma(c)$ can be thought of as a virtual cost function.[8] For most of the analysis, we simplify notation by assuming that $\underline{c} = \underline{v}$ and $\overline{c} = \overline{v}$, an assumption that turns out not to be restrictive.[9]

A sequence of fee functions $\boldsymbol{\omega} = (\omega_t)_{t=0}^{\infty}$ with $\omega_t(\breve{p})$ specifies the amount the seller has to pay to the intermediary when a transaction occurs in period t at the transaction price $\breve{p}$. To fix ideas, we assume that the transaction price is determined by an English auction with reserve price p_t set by the seller. It may be literally the case that the seller uses an English auction or the English auction may serve as a model for the way bargaining between the seller and buyers unfolds.[10] The game ends in period t when a buyer bids

[7]Formally, the respective distributions and supports are given by $G(c) := G_0((c - K^S)/\hat{K}^S)$ with support $[\underline{c}, \overline{c}]$ and $F(v) := F_0((v - K^B)/\hat{K}^B)$ with support $[\underline{v}, \overline{v}]$, where $\underline{c} := \hat{K}^S \underline{c}_0 + K^S$, $\overline{c} := \hat{K}^S \overline{c}_0 + K^S$, $\underline{v} := \hat{K}^B \underline{v}_0 + K^B$, and $\overline{v} := \hat{K}^B \overline{v}_0 + K^B$.

[8]Interpreting $G(p)$ and $1 - F(p)$ as expected quantities supplied and demanded, $\Phi(p)$ and $\Gamma(p)$ have the interpretation of marginal revenue and marginal cost functions (see Bulow and Roberts, 1989): Setting $q = 1 - F(p)$, the revenue when selling q is $F^{-1}(1-q)q$. Consequently, marginal revenue is $[F^{-1}(1-q)q]'_{p=F^{-1}(1-q)} = \Phi(p)$. Analogous reasoning applies to $\Gamma(p)$.

[9]If one starts out with $\underline{c}_0 = \underline{v}_0$ and $\overline{c}_0 = \overline{v}_0$, then after taking into account the additional transaction costs, one ends up with $\underline{c} \leq \underline{v}$ and $\overline{c} \leq \overline{v}$. Sellers with costs $c > \overline{v}$ cannot find a buyer with whom they have positive gains from trade and hence can be ignored. Similarly, buyers with valuations $v < \underline{c}$ cannot find a seller with whom they have positive gains from trade and hence can be ignored. Therefore, we only need to consider sellers and buyers in the interval $[\underline{c}, \overline{v}]$ and truncate and rescale G and F to have this common support.

[10]There are many setups that are formally equivalent. For example, given that buyers have dominant

higher than p_t.[11] The seller is not allowed to recall buyers after the period in which they arrived.[12] We assume full commitment by the intermediary throughout the paper.

The above specification is sufficiently general to include a number of setups of applied interest. To model auction platforms and auction houses such as eBay, Sotheby's, or Christie's, one may consider a one-shot setup ($\delta = 0$) and a binomial distribution of buyers ($\overline{B} > 1$) who literally participate in an English auction. To capture Amazon's third-party sellers platform, one can think of the third-party seller as offering the good to a potential buyer at a fixed price in a single period ($\delta = 0$). The single buyer corresponds to setting $\overline{B} = 1$, in which case the English auction with a reserve price reduces to a take-it-or-leave-it price offer equal to the reserve. To model real-estate brokerage, one can assume that there is a seller who offers his house in multiple periods ($\delta > 0$) to buyers whose number is Poisson distributed in every period,[13] with bargaining modeled as an English auction (in real-estate transactions bargaining is typically intermediated by the broker who keeps buyers informed about the highest standing offer, so that the ensuing bargaining game is equivalent to an English auction). In all these cases, intermediaries raise revenues by charging transaction fees.

3 Optimal Transaction Fees

For a given ω_t, the seller's expected net revenue $R_{\omega_t}(p_t)$ in period t conditional on a transaction occurring and given reserve p_t is

$$R_{\omega_t}(p_t) = \frac{(p_t - \omega_t(p_t))(F_{(2)}(p_t) - F_{(1)}(p_t)) + \int_{p_t}^{\overline{v}} (\breve{p} - \omega_t(\breve{p}))dF_{(2)}(\breve{p})}{1 - F_{(1)}(p_t)},$$

strategies, it does not matter whether the reserve price is public or remains private information of the seller. The bargaining may also be such that the seller keeps rejecting bids that are below the maximum of his reserve and the highest standing bid by any buyer, allowing rejected buyers to revise their bids upwards. As briefly discussed in Section 5.3, it is also immaterial whether the auction format is an English auction or a first-price auction, provided fees are linear and the reserve price is known by the time buyers submit their bids.

[11] If $1 - \delta$ is interpreted as the probability that the seller drops out from one period to the next, the game can also end when the seller drops out.

[12] As shown by Riley and Zeckhauser (1983), this assumption is without loss of generality with a commonly known distribution F when the seller can commit to an optimal strategy and when one buyer enters in every period.

[13] Recall that a Poisson arrival rate is the limit of the binomial when $\overline{B} \to \infty$ and the expected number of buyers $\tilde{\pi}\overline{B}$ is kept constant.

by standard arguments from auction theory (see e.g. Krishna, 2002), where $F_{(1)}(v) := \sum_{B=0}^{\infty} \pi_B F(v)^B$ and $F_{(2)}(v) := F_{(1)}(v) + (1 - F(v)) \sum_{B=1}^{\infty} \pi_B B F(v)^{B-1}$ are, respectively the unconditional distribution of the highest and the second-highest valuation.[14] Consequently, the maximization problem of a seller of type c given $\boldsymbol{\omega}$ is to choose a sequence of prices $\boldsymbol{p} = (p_t)_{t=0}^{\infty}$ to maximize his discounted expected profit

$$W_S(c, \boldsymbol{p}, \boldsymbol{\omega}) := \sum_{t=0}^{\infty} (R_{\omega_t}(p_t) - c)(1 - F_{(1)}(p_t)) \prod_{\tau=0}^{t-1} \delta F_{(1)}(p_\tau),$$

where we use the convention of defining $\prod_{\tau=0}^{t-1} \delta F_{(1)}(p_\tau) = 1$ for $t = 0$. Let $\boldsymbol{P}(c) = (P_t(c))_{t=0}^{\infty}$ be the (or a) maximizer of $W_S(c, \boldsymbol{p}, \boldsymbol{\omega})$, whose dependence on the sequence of fees is kept implicit for ease of notation.

Given ω_t, the intermediary's expected revenue in period t when facing a seller of type c who sets the reserve price $p_t = P_t(c)$ is $\omega_t(p_t)(F_{(2)}(p_t) - F_{(1)}(p_t)) + \int_{p_t}^{\overline{v}} \omega_t(\breve{p}) dF_{(2)}(\breve{p})$. Therefore, the intermediary's discounted expected profit from a seller of type c given $\boldsymbol{p}$ and $\boldsymbol{\omega}$ is

$$W_I(c, \boldsymbol{p}, \boldsymbol{\omega}) := \sum_{t=0}^{\infty} \left(\omega_t(P_t(c))(F_{(2)}(P_t(c)) - F_{(1)}(P_t(c))) + \int_{P_t(c)}^{\overline{v}} \omega_t(\breve{p}) dF_{(2)}(\breve{p}) \right) \prod_{\tau=0}^{t-1} \delta F_{(1)}(P_\tau(c)).$$

We assume that the fees $\boldsymbol{\omega}$ are chosen to maximize a weighted average of the expectation of the intermediary's profit W_I and of the joint surplus of the intermediary and the seller $W_I + W_S$

$$W(\alpha, \boldsymbol{\omega}) := E_{c\sim G}[\alpha W_I(c, \boldsymbol{P}(c), \boldsymbol{\omega}) + (1 - \alpha)(W_I(c, \boldsymbol{P}(c), \boldsymbol{\omega}) + W_S(c, \boldsymbol{P}(c), \boldsymbol{\omega}))], \quad (1)$$

where $\alpha \in [0, 1]$ is a parameter measuring the intermediary's bargaining power. It can also be interpreted as a measure of competition between brokers for sellers, with $\alpha = 0$ corresponding to perfect competition, $\alpha = 1$ corresponding to monopoly power (or perfect collusion by intermediaries), and $\alpha \in (0, 1)$ can be seen as capturing imperfect competition or imperfect collusion in a reduced form.[15] The resulting fee structure can

[14]The above expression can be obtained by observing that $1 - F_{(1)}(p_t)$ is the probability that a transaction occurs, the probability that the transaction price is equal to the reserve is $F_{(2)}(p_t) - F_{(1)}(p_t)$, and the distribution of transaction prices above the reserve is $\breve{p} \sim F_{(2)}$.

[15]One way to think about perfect competition is that there are multiple sellers offering goods which are not substitutable. Each seller is matched with one intermediary. Competition drives down profits

then be seen as the outcome of bargaining between the intermediary and the seller. As shown below, $\alpha = 0$ implies that the fees are 0 for all prices. Observe also that the objective function in (1) depends on $\boldsymbol{\omega}$ directly and also indirectly via the seller's pricing behavior $\boldsymbol{P}(c)$, which depends on $\boldsymbol{\omega}$.[16]

The assumption that the intermediary and the seller bargain over the division (and size) of their joint surplus captures the notion that in many markets of interest, in particular in real-estate markets, sellers typically sign long-term exclusive dealership contracts with brokers. According to our modeling choice, sellers who are more patient than others would be characterized by larger opportunity costs of selling. Although space constraints refrain us from so doing, the setup can be extended to allow for sellers who have heterogenous deadlines and discount in non-stationary ways. Below we will show that our assumptions regarding the menu of mechanisms that can be chosen are without loss of generality. We will show that it is optimal to choose a second-price auction with a reserve price set by the seller with an appropriately chosen fee structure, and we discuss conditions under which the results extend to first-price auctions. The assumptions that the environment is stationary and that F and G have the same support and exhibit monotone virtual valuation and cost functions are imposed for expositional simplicity as they do not affect the key insights from our analysis.[17]

It is important to keep in mind that optimality in our context means optimal pricing by an intermediary to extract rents (for $\alpha > 0$). It does not imply that the fees charged are socially optimal: as usual when dealing with optimal pricing by a firm that has

to zero ($\alpha = 0$). For imperfect competition, $\alpha \in (0, 1)$ can be thought of as a conduct parameter that captures in a reduced form any reason that the intermediaries might extract positive rents, such as service differentiation between intermediaries or capacity constraints. See e.g. Brander and Zhang (1990) and the references therein for the empirical industrial organization literature using the conduct parameter approach and e.g. Corts (1999) for a discussion of the limitations of the conduct parameter approach. One would think of perfect collusion of multiple intermediaries as the intermediaries colluding not to try to attract the sellers of other intermediaries by lowering their fees. Collusion may be imperfect ($\alpha \in (0, 1)$) due to the colluding intermediaries trying to avoid detection by anti-trust authorities or collusion being easier to sustain in a repeated game if the collusive fees are lower.

[16]Maximizing $W(\alpha, \boldsymbol{\omega})$ is equivalent to maximizing the weighted sum $E_c[\alpha_0 W_I(c, \boldsymbol{P}(c), \boldsymbol{\omega}) + (1 - \alpha_0) W_S(c, \boldsymbol{P}(c), \boldsymbol{\omega})]$ with $\alpha_0 = 1/(2 - \alpha)$.

[17]In a previous version of our paper (Loertscher and Niedermayer, 2012), we have derived results for the cases when these assumptions do not hold. The results gave essentially the same economic insights, but the notation was far more tedious.

market power and extracts rents, the pricing is not socially optimal.[18]

3.1 Seller Behavior

Given a sequence of fee functions $\boldsymbol{\omega} = (\omega_t)_{t=0}^{\infty}$ the seller will choose a sequence of reserve prices $\mathbf{p} = (p_t)_{t=0}^{\infty}$ to maximize the expected net present value of his profits. In general, this maximization problem is complex because the fees could be non-stationary and the implied profit function of the seller could fail to be quasi-concave, so that the first-order condition would not be sufficient. However, we will later show that even with an arbitrary non-stationary mechanism one could not do better than one can by charging fees which are stationary and imply a quasi-concave profit function for the seller. Therefore, to reduce the notational burden, we will focus on stationary fees and reserve prices, that is $\omega_t = \omega$ and $p_t = p$ for all t, and use the first-order condition for maximization.[19]

Given stationary fees $\boldsymbol{\omega}$ and stationary prices $\boldsymbol{p}$, the seller's utility becomes

$$W_S(c, \boldsymbol{p}, \boldsymbol{\omega}) = (R_\omega(p) - c)(1 - F_\infty(p)),$$

where

$$1 - F_\infty(p) := (1 - F_{(1)}(p)) \left(\sum_{t=0}^{\infty} \delta^t F_{(1)}(p)^t \right) = \frac{1 - F_{(1)}(p)}{1 - \delta F_{(1)}(p)} \tag{2}$$

is the *ultimate probability of selling* (which is inspired by Satterthwaite and Shneyerov (2007, 2008)'s notion of the "ultimate discounted probability of trade").[20] Let

$$\Phi_\omega(p) := p - \omega(p) - (1 - \omega'(p)) \frac{1 - F(p)}{f(p)}$$

[18]We do not deal with the question of social optimality, because of different controversial aspects of many intermediaries which are orthogonal to the research question (optimal pricing) of this paper. Intermediaries may extract rents to cover fixed costs of operation, which is second-best efficient if one does not want government subsidized (or even government run) intermediaries. There is some controversy surrounding private intermediaries, e.g. the International Labor Organization demanded a ban of private fee-charging labor market intermediaries in its C96 convention (Fee-Charging Employment Agencies Convention (Revised), 1949). Instead, they demanded public employment agencies. Even if one agrees on having private intermediaries, one may be skeptical of an intermediary's ability of extracting rents, since an intermediary's market power may be due to collusion. For example, there is an allegation of collusion for real estate agents and a conviction for collusion of Sotheby's and Christie's.

[19]By using standard techniques it is possible to extend the analysis to non-optimal fees which imply a non-stationary non-quasi-concave problem.

[20]If one interprets the discount factor as the probability of drop-out, $1 - F_\infty$ is the probability of selling taking into account that one might die with probability $1 - \delta$ in every period.

be the net virtual valuation associated with the stationary fee ω, and define

$$\tilde{\Phi}_\omega(p) := \overline{v} - \omega(\overline{v}) - \int_p^{\overline{v}} \frac{1 - \delta F_{(1)}(v)}{1 - \delta} \Phi'_\omega(v) dv.$$

The function $\tilde{\Phi}_\omega(p)$ is monotone and thus invertible if $\Phi_\omega(p)$ is monotone.

The seller chooses the reserve p to maximize W_S. The following proposition gives the solution to this maximization problem.

Proposition 1. *Given a stationary fee ω that implies a monotone net virtual valuation $\Phi_\omega(p)$, the optimal price set by a seller with cost c is $P(c) = \tilde{\Phi}_\omega^{-1}(c)$ in every period.*

As mentioned, we will show that the optimal fee is such that the seller's profit function is quasi-concave, which is equivalent to an increasing Φ_ω. One can show that Φ_ω (and hence also $\tilde{\Phi}_\omega$) is increasing, provided Φ is increasing and the fee ω is linear with slope less than 1, which proves helpful in empirical applications. For notational ease, we let $\tilde{\Phi}(v) := \tilde{\Phi}_0(v)$ and $R(p) := R_0(p)$. Note that $\tilde{\Phi}_\omega$ can be interpreted as the net dynamic virtual valuation function and satisfies $\tilde{\Phi}_\omega(\overline{v}) = \overline{v}$. In a static setup ($\delta = 0$), $\tilde{\Phi}_\omega$ simplifies to the net virtual valuation Φ_ω. If the fee is zero ($\omega(p) = 0$ for all p), it further simplifies to the virtual valuation function Φ.

Because of their widespread use, percentage fees $\omega(p) = bp$ with $b \in [0, 1]$ are of particular interest. Abusing notation, we write $\Phi_b(v) := \Phi_\omega(v)|_{\omega(p)=bp}$ and $\tilde{\Phi}_b(v) := \tilde{\Phi}_\omega(v)|_{\omega(p)=bp}$. We have $\Phi_b(v) = (1 - b)\Phi(v)$ and $\tilde{\Phi}_b(p) = (1 - b)\tilde{\Phi}(p)$. Consequently, for a percentage fee b Proposition 1 implies that the optimal price set by a seller with cost c is

$$P(c) = \tilde{\Phi}^{-1}(c/(1 - b)). \tag{3}$$

The intuition for the above equation is that $\tilde{\Phi}^{-1}(x)$ is the profit maximizing price set by a monopolist with (gross) cost x. The net costs c of the seller (before fees) are $1 - b$ times the gross costs, hence the gross costs are $x = c/(1 - b)$.

Proposition 1 relates the seller's optimal reserve price $P(c)$ to the fee function ω. Once we know ω, we know $P(c)$ provided ω is monotone. Because the good will be sold to the buyer with the highest value in the first period in which this value exceeds $P(c)$, the search for the optimal fee function can be separated into two steps. First, find the pricing function that is jointly optimal for the intermediary and the seller. Second, find the fee function that induces the seller to set the optimal price. A priori it is not clear whether such a fee function exists. A key insight of our paper is to show that it does.

3.2 Optimality of Fees

The problem of maximizing $W(\alpha, \boldsymbol{\omega})$ over $\boldsymbol{\omega}$ is tedious. Moreover, even if one succeeds in solving this problem one will not know whether the use of fee setting is without loss of generality. In Appendix A.2 we set up and solve the general mechanism design problem for our model, without imposing any constraints on the mechanism other than incentive compatibility[21] and individual rationality[22]. Mechanisms that solve this problem are called optimal. We show there that the focus on direct mechanisms – these are mechanisms that ask each agent to report his type upon arrival, provided the seller is still in the game – is without loss of generality and that revenue equivalence holds. That is, once the allocation rule is determined, the interim expected payoff of every agent of every type is determined by the allocation rule up to an additive constant, which in the optimal mechanism is set equal to zero because the individual rationality constraints will optimally bind.

Let $\Gamma_\alpha(c) := \alpha\Gamma(c) + (1-\alpha)c$ be the weighted average of the seller's virtual cost $\Gamma(c)$ and his type c. The key result from the mechanism design analysis is the following:

Lemma 1. *In any optimal mechanism, the good is sold, to the buyer with the highest valuation present in that period, in the earliest period t for which*

$$\max_{i \in \mathcal{B}_t} \tilde{\Phi}(v_i) \geq \Gamma_\alpha(c),$$

where $\mathcal{B}_t$ is the set of buyers who arrived in period t, and the expected payoff of every buyer of type $\underline{v}$ and of the seller of type $\overline{c}$ is 0.

While the result is intuitive, the proof is surprisingly involved. The reason why one cannot use standard mechanism design techniques is that potential future buyers have

[21] No participant has an incentive to choose an option that is meant for a participant of another type.
[22] Participants are willing to choose the option offered to them rather than the outside option.

not yet arrived, so they cannot be asked to reveal their types in the beginning. Further, it is not obvious how to compare probabilities of transactions and revenues from transactions today with probabilities and revenues in the future, because of discounting. To get around these difficulties, we use the concept of the *ultimate (discounted) probability of trade* introduced in Satterthwaite and Shneyerov (2007, 2008). Since on top of the dynamic bargaining game with private information considered in Satterthwaite and Shneyerov (2007, 2008), we also have a mechanism design problem, we need to introduce an additional concept, the *ultimate conditional expected revenue*. These concepts are described in more detail in the proof in the Appendix.[23]

Lemma 1 generalizes Theorem 3 of Myerson and Satterthwaite (1983) to our dynamic setting with multiple buyers using the concept of the dynamic virtual valuation. It is based on the insight that in any optimal mechanism, the good goes to the buyer with the highest value in any given period if this value is above some threshold, and stays with the seller otherwise. Lemma 1 adds to this the insight that the good goes to the buyer with the highest dynamic virtual valuation, appropriately defined, provided it exceeds $\Gamma_\alpha(c)$. Intuitively, Myerson and Satterthwaite (1983) find in a one-shot setup with one buyer and one seller that the good is transferred whenever $\Phi(v)$ exceeds $\Gamma(c)$. In the language of optimal monopoly pricing, this means that the good changes hands whenever marginal revenue is larger than marginal cost (see footnote 8 for why virtual types can be interpreted as marginal revenue and marginal cost). In our setup, the dynamic virtual valuation $\tilde{\Phi}(v)$ has to be used to adjust for the option value of future trade and the weighted virtual cost $\Gamma_\alpha(c)$ to account for the weight on the seller's utility.

In light of the remarks after Proposition 1, Lemma 1 answers the first question: it derives the optimal allocation rule. The optimal allocation rule can be implemented via fee setting if a seller of type c can be induced to set the reserve price

$$P^*(c) := \tilde{\Phi}^{-1}(\Gamma_\alpha(c)) \tag{4}$$

in every period in which he is active. If this is possible, then bidding in the English auction will ensure that the object goes to the buyer with the highest virtual valuation

[23]The ultimate conditional expected revenue sounds similar to, but is distinctively different from the expected net present value of the revenue.

while the reserve price $P^*(c)$ ensures that trade only takes place if this virtual value, that is $\tilde{\Phi}$, exceeds $\Gamma_\alpha(c)$. The discounted probability that a seller of type c who always sets the price $P^*(c)$ ever sells is $1 - F_\infty(P^*(c))$. A seller with cost $\Gamma_\alpha^{-1}(\overline{v})$ should optimally set the price $\overline{v}$ and never trade. By a standard revenue equivalence argument,[24] the expected discounted payoff $V(c)$ of a seller of type $c \in [\underline{c}, \Gamma_\alpha^{-1}(\overline{v})]$ who always sets the price $P^*(c)$ is pinned down by the allocation rule and given by

$$V(c) = \int_c^{\Gamma_\alpha^{-1}(\overline{v})} (1 - F_\infty(P^*(y)))dy.$$

With this in hand, we can now describe the optimal transaction fees $\omega_t(\breve{p})$.

Proposition 2. *The optimal transaction fees implement the optimal mechanism whose allocation rule is described in Lemma 1 and are such that for all* $t = 0, 1, ..$

$$\omega_t(p) = \omega(p) := p - \frac{\int_p^{\overline{v}} \left[\Gamma_\alpha^{-1}(\tilde{\Phi}(v)) + \delta V(\Gamma_\alpha^{-1}(\tilde{\Phi}(v)))\right] f(v)dv}{1 - F(p)}. \tag{5}$$

The proof that this fee induces the seller of type c to set the price $P^*(c)$ in every period is surprisingly simple. By the one-period-deviation principle, we can confine attention to a deviation by the seller in the present period to some reserve price p and assume that the seller sets the price $P^*(c)$ in every period after that, whereby he gets $\delta V(c)$. The expected payoff from so doing given the fee $\omega(p)$ defined in (5) is $(R_{\omega_t}(p) - c)(1 - F_1(p))$ in the period of deviation and $F_1(p)\delta V(c)$ afterwards, which can be rearranged to

$$(p - \omega(p))[F_{(2)}(p) - F_{(1)}(p)] + \int_p^{\overline{v}} [y - \omega(y) f_{(2)}(y)dy + F_{(1)}(p)[c + \delta V(c)].$$

The first-order condition for a maximum is

$$0 = f_{(1)}(p) \left[-\Gamma_\alpha^{-1}(\tilde{\Phi}(p)) - \delta V(\Gamma_\alpha^{-1}(\tilde{\Phi}(p))) + c + \delta V(c)\right],$$

which follows after substituting $\omega(\cdot)$ and cancelling terms (in particular, using the fact that $F_{(2)}(p) - F_{(1)}(p) = f_{(1)}(p)(1 - F(p))/f(p)$). The first-order condition is satisfied at $p = P^*(c)$ and because the term in brackets decreases in p, it follows that the objective

[24] See e.g. Myerson (1981).

function is quasi-concave, implying that the first-order condition is sufficient for a maximum. Importantly, Proposition 2 implies that fee setting with the fee given in (5) is optimal in the domain of all incentive compatible, individually rational mechanisms.

As an illustration, consider the static setup by setting $\delta = 0$, which has been studied extensively in the literature. With $\pi_1 = 1$, we have one buyer with certainty, and if we set $\alpha = 1$, our fee-setting mechanism implements the broker-optimal mechanism derived by Myerson and Satterthwaite (1983). For the example in Myerson and Satterthwaite (1983) (F and G uniform on $[0, 1]$), the broker-optimal fee is 50%.[25] It is easy to check that the broker's expected revenue is 1/24, which is the same as for the direct mechanism derived in Myerson and Satterthwaite (1983). For any fixed number of buyers $\overline{B}$ (i.e. $\pi_{\overline{B}} = 1$) and any distribution F satisfying regularity, our fee-setting mechanism specializes to the optimal auction with reserve $P^*(c) = \Phi^{-1}(c)$ of Myerson (1981) if all the weight is on the seller's welfare ($\alpha = 0$).

4 Optimal Linear Fees in Thin Markets

In principle, the optimal fee schedule can be a complicated non-linear function. Empirically, however, simple, linear fees are often used. Linear fees are particularly prevalent in thin markets such as real-estate markets and high-skill labor markets, where typically only a small percentage of potential sellers is active in the market. As an example, typically less than 5% of home owners offer their property for sale at a given point of time. Amazon's fees for third-party sellers of most types of goods (including books, consumer electronics, and personal computers) are another case in point. We now show how additional transaction costs whose presence induces only the most motivated traders to participate imply that the optimal fees will be asymptotically linear. We do so by applying results from extreme value theory to markets with fee setting.

There are various possible and mutually non-exclusive sources of such costs. The cost of physical relocation – of moving or shipping – is one that is due to exogenous costs.

[25] The uniform is a special case of a mirrored Generalized Pareto distribution $G(c) = c^{\sigma}$ for $c \in [0, 1]$ with $\sigma > 0$, which yields as the broker-optimal fee of $\omega(p) = p/(1 + \sigma)$ for the static setup for any F and $\tilde{\pi}$.

In dynamic models, such transaction costs may also arise endogenously from the agents' opportunity costs of future trade, which in any given period makes agents less inclined to trade. To fix ideas, we will focus on the case of exogenous transaction costs, and we will assume that after the realization of their types, and knowing the transaction costs, agents can decide whether they want to participate in the market. Later on, we will discuss in more detail microfoundations for such transaction costs.

To capture the notion that only a small fraction of potential traders are active, we introduce increasing transaction costs. For simplicity, we normalize the supports of the primitive distributions F_0 and G_0 from which buyers and sellers draw their primitive types v_0 and c_0 to $[0, 1]$. We study a sequence of economies characterized by transaction costs $\mathbf{K}_j := (K_j^S, \hat{K}_j^S, K_j^B, \hat{K}_j^B)$ and focus on the limit of this sequence, indexed by $j \geq 0$, as the cost becomes large with $\mathbf{K}_0 = (0, 1, 0, 1)$. The distribution of effective costs $c = K_j^S + \hat{K}_j^S c_0$ is $G_j(c) = G_0((c - K_j^S)/\hat{K}_j^S)$ and the distribution of effective valuation $v = K_j^B + \hat{K}_j^B v_0$ is $F_j((v - K_j^B)/\hat{K}_j^B)$. Our previous analysis directly applies by replacing F by F_j and G by G_j.

There are many different ways in which transaction costs may reduce the fraction of active traders. One example are moving costs for a buyer of a property, which are additive ($K_j^B > 0$, $\hat{K}_j^B = 0$). Another example is an option value x for the buyer of real estate, which may be due to the possibility of buying another property or the possibility of renting, such that the buyer's willingness to pay is $v = \lambda_j x + (1 - \lambda_j)v_0$, where λ_j is a weight put on the option value that will be discussed later. For this example $K_j^B = \lambda_j x$ and $\hat{K}_j^B = 1 - \lambda_j$. The same applies for the seller's transaction costs K_j^S and $\hat{K}_j^S$. There are many different combinations of changes of K_j^B, $\hat{K}_j^B$, K_j^S, and $\hat{K}_j^S$ that lead to a decrease of the fraction of active traders. However, it is not necessary to go through all combinations. Instead, one can greatly simplify the analysis by introducing two variables u_j^B and u_j^S whose decrease implies a decrease of the fraction of active traders. Therefore, we defer providing microfoundations of different changes of $\mathbf{K}_j$ to Section 5.1 and turn to the variables u_j^B and u_j^S in the following.

Denote the implied supports with $[\underline{c}_j, \overline{c}_j]$ and $[\underline{v}_j, \overline{v}_j]$, respectively. In the following, it will be useful to think of the *relevant range* $[\underline{c}_j, \overline{v}_j]$ in which the two supports overlap. It is also useful to define the ratio of the length of the relevant range to the length of the seller's support $u_j^S := (\overline{v}_j - \underline{c}_j)/(\overline{c}_j - \underline{c}_j)$. Analogously, define $u_j^B := (\overline{v}_j - \underline{c}_j)/(\overline{v}_j - \underline{v}_j)$ for the buyer. Since there is a one-to-one mapping between the set of parameters $(\underline{c}_j, u_j^S, \overline{v}_j, u_j^B)$ and $\boldsymbol{K}_j$, we can write the following analysis in terms of $(\underline{c}_j, u_j^S, \overline{v}_j, u_j^B)$.

Since sellers with $c > \overline{v}_j$ trade with probability 0, a seller participates if and only if $c \leq \overline{v}_j$, which is equivalent to $c_0 \leq u_j^S$. Therefore, the mass of active sellers is $G_0(u_j^S)$. Analogously, the mass of active buyers is $1 - F_0(1 - u_j^B)$.

The analysis simplifies by normalizing $\tilde{c} := (c_0 - \underline{c}_j)/(\overline{v}_j - \underline{c}_j)$, $\tilde{v} := (v_0 - \underline{c}_j)/(\overline{v}_j - \underline{c}_j)$, and $\tilde{p} := (p - \underline{c}_j)/(\overline{v}_j - \underline{c}_j)$. The distributions of the normalized effective cost $\tilde{c}$ and the normalized effective valuation $\tilde{v}$, truncated to $[\underline{c}_j, \overline{v}_j]$ and denoted, respectively, $\tilde{G}_j$ and $\tilde{F}_j$, are then given as

$$\tilde{G}_j(\tilde{c}) := \frac{G_0(u_j^S \tilde{c})}{G_0(u_j^S)} \quad \text{and} \quad \tilde{F}_j(\tilde{v}) := 1 - \frac{1 - F_0(1 - u_j^B(1 - \tilde{v}))}{1 - F_0(1 - u_j^B)},$$

with the normalized fee defined as $\tilde{\omega}_j(\tilde{p}) = \omega(p)/(\overline{v}_j - \underline{c}_j)$.[26]

The following proposition relies on Extreme Value Theory, which states that the upper tail of any distribution converges to a Generalized Pareto distribution as one moves the truncation point closer to the upper bound of the support, as long as the distribution satisfies some weak regularity assumptions (see Appendix C for more details on extreme value theory and also for a version of the theory with an infinite upper bound of the support). These regularity assumptions can be shown to be satisfied in our setup. Further, analogous mirror image results hold with regards to the lower bound of the support. The proposition also shows the relation between Extreme Value Theory and linear fees.

Proposition 3. *Let the shifting constants $\underline{c}_j$, $\overline{v}_j$ be arbitrary sequences satisfying $\underline{c}_j < \overline{v}_j$ for all j. Let the ratios of the relevant ranges u_j^S and u_j^B be sequences that go to 0 as j goes to infinity. Then, as $j \to \infty$,*

(i) the buyers' and the seller's normalized distributions converge to Generalized Pareto and mirrored Generalized Pareto distributions, respectively: $\lim_{j\to\infty} \tilde{F}_j(\tilde{v}) = \tilde{F}^*(\tilde{v}) :=$

[26]Despite notational similarities, the distribution $\tilde{F}_j$ has no relation to the dynamic virtual valuation $\tilde{\Phi}$ introduced after Proposition 1 above. We will not use virtual valuations associated with $\tilde{F}_j$.

$1-(1-\tilde{v})^\beta$ *and* $\lim_{j\to\infty}\tilde{G}_j(\tilde{c})=\tilde{G}^*(\tilde{c}):=\tilde{c}^\sigma$ *for some constants* β *and* σ.

(ii) the normalized fee $\tilde{\omega}_j(\tilde{p})$ *converges to* $\alpha\tilde{p}/(\alpha+\sigma)$*, that is:*

$$\lim_{j\to\infty}\tilde{\omega}_j(\tilde{p})=\frac{\alpha}{\alpha+\sigma}\tilde{p}. \tag{6}$$

The constants β and σ in the Proposition are referred to as the Pareto tail indices in the statistics literature (see Section C for some of the references) and have been observed to be remarkably invariant over time for a number of applications.

We first provide an intuition for part (i) of the Proposition. Convergence of the distribution is immediate when the primitive distributions G_0 and F_0 are (mirrored) Generalized Pareto distributions on $[0,1]$, that is if $G_0(c_0)=c_0^\sigma$ for $\sigma>0$ and $F_0(v_0)=1-(1-v_0)^\beta$ with $\beta>0$, for this implies $\tilde{G}_j(\tilde{c})=G_0(u_j^S\tilde{c})/G_0(u_j^S)=\tilde{c}^\sigma$ and $1-\tilde{F}_j(\tilde{v})=(1-F_0(1-u_j^B(1-\tilde{v}))/(1-F_0(1-u_j^B))=(1-\tilde{v})^\beta$ for all j. In other words, $\tilde{G}_j$ and $\tilde{F}_j$ do not change with j. This is, of course, the well-known property of Pareto distributions that they are invariant to truncation.

Next, we discuss the correct interpretation of the asymptotic results when one starts away from the limiting distribution. This is important since it is tempting to interpret asymptotic results as only being applicable in one of two cases: either if one assumes very particular functional forms for distributions that are close to the limiting distribution or if one is exactly in the limit. However, this would be a misinterpretation. While for often used asymptotic results such as the central limit theorem the correct interpretation seems obvious (the misinterpretation is well understood to be a false dichotomy),[27] for other asymptotic results, such as extreme value theory, the correct interpretation does not appear to be that obvious.[28]

[27]For the central limit theorem, such a misinterpretation would mean that the theorem is only applicable in one of two cases: (i) the distribution of a variable has a peculiar functional form that is very close to a normal distribution to start with or (ii) one has to be (almost) exactly in the limit, which means taking the average of an infinite number of random draws. One would then *erroneously* believe that either one has to make an overly restrictive assumption on functional forms (case (i)) or that the variance of the average is (almost exactly) zero (since we are taking the average of an infinite number of random draws, case (ii)). However, the distinction of cases (i) and (ii) is a false dichotomy: the applicability of the central limit theorem is due to the middle ground between cases (i) and (ii).

[28]For extreme value theory, the misinterpretation is that asymptotic results apply in only one of two cases: (i) the distribution is close to Generalized Pareto to begin with or (ii) the mass in the tail of the truncated distribution is (close to) zero. Again, a false dichotomy.

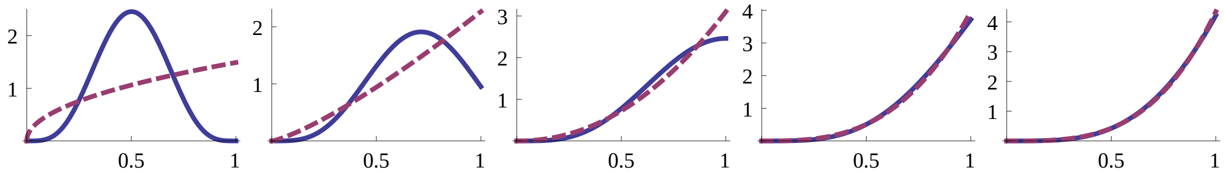
We take two strategies to avoid potential misinterpretations. First, we provide numerical results that illustrate that even when starting with a distribution that is far from the limiting distribution and when not going too close to the limit, the results of Extreme Value Theory are already a good approximation. Second, in our empirical companion paper (Loertscher and Niedermayer, 2017), we provide an empirical analysis that shows that, for the empirically estimated distributions, Extreme Value Theory is a good approximation.

Consider the following numerical example in which G_0 is a Beta-distribution with density $g_0(c_0) \propto c_0^4(1-c_0)^4$ on the support $[0,1]$. As panel (a) of Figure 1 shows, this is far from a Generalized Pareto distribution. The figure also displays the distribution conditional on $c_0 \in [0,u]$ for $u \in \{1, 0.7, 0.5, 0.3, 0.2\}$. Moving the truncation point u downwards brings the density of the conditional distribution closer to the density of a mirrored Generalized Pareto distribution. Of course, the statement of Extreme Value Theory holds in the limit as is always the case for asymptotic results. However, as usual in statistics, one should interpret asymptotically founded results as providing good approximations away from the limit. For example, panel (d) in Figure 1 depicts the case when still ten percent of all seller types are active. The distribution is already very well approximated by a mirrored Generalized Pareto distribution. This is even more so in panel (e), when two percent of sellers are active. In this case, the overlap is almost perfect. One may wonder how close we are at the limit in practice. This question can only be answered empirically, which we do in Loertscher and Niedermayer (2017).

The intuition for part (ii) of Proposition 3 is most easily gleaned by specializing to a static setup (i.e. $\delta = 0$) and assuming that G_0 is a mirrored Generalized Pareto distribution, that is $G_0(c_0) = c_0^\sigma$ for $c_0 \in [0,1]$. This implies that the virtual cost function is linear, that is, $\Gamma_{\alpha,0}(c_0) := c_0 + \alpha G_0(c_0)/g_0(c_0) = c_0(1+\alpha/\sigma)$.[29] The optimal fee can thus be written as

$$\omega(p) = p - E_{v_0 \sim F_0}[\Gamma_{\alpha,0}^{-1}(\Phi_0(v_0))|v_0 \geq p]. \tag{7}$$

[29] By adjusting the support appropriately, for any element j in the sequence a similar analysis applies and delivers completely analogous results if G_0 is a mirrored Generalized Pareto distribution because of the truncation invariance of the virtual cost and of these distributions.



(a) $u = 1$ (b) $u = 0.7$ (c) $u = 0.5$ (d) $u = 0.3$ (e) $u = 0.2$

Figure 1: Density of truncated, rescaled distribution $G_u(c) = G_0(\underline{c}+u(c-\underline{c}))/G_0(\underline{c}+u(\bar{c}-\underline{c}))$ for $u \in \{1, 0.7, 0.5, 0.3, 0.2\}$ for a Beta distribution with support $[0,1]$ and density $g_0(c) \propto c^4(1-c)^4$ (solid line) compared to an approximating mirrored Generalized Pareto density with support $[0,1]$ (dashed). Masses in the relevant range are (a) $G(1) = 1$, (b) $G_0(0.7) = 0.9$, (c) $G_0(0.5) = 0.5$, (d) $G_0(0.3) = 0.1$, (e) $G_0(0.2) = 0.02$. As the mass decreases, the distribution converges to the approximating Pareto distribution and the approximating Pareto distribution converges to the limiting Pareto distribution.

Because it is linear, one can pull $\Gamma_{\alpha,0}^{-1}$ outside the expectation, and because $E_{v_0 \sim F_0}[\Phi_0(v_0)|v_0 \geq p] = p$, one obtains $\omega(p) = p - \Gamma_{\alpha,0}^{-1}(p) = p\alpha/(\alpha+\sigma)$.[30]

The reasoning in the above paragraph and part (i) of Proposition 3 do not yet constitute a proof of part (ii) of the proposition since one still has to establish that the transformation from distributions to fees is continuous. The proof is quite lengthy, which may not come as a surprise given the complexity of expression in (5) and the continuation value $V(\cdot)$. We relegate the proof to Appendix A and provide an illustration here, based on the same numerical example used for Figure 1. Figure 2 shows that the optimal fee moves closer to a linear fee as the transaction costs increase. The mass of traders $G_0(u)$ and $1 - F_0(1-u)$ does not have to be very close to 0 for the optimal fees to be close to linear. In Figure 2 (d) and (e) the optimal fee is already well approximated by a linear fee with the masses of traders being ten percent and two percent, respectively, and the length of the support u being 30 percent and 20 percent of the length of the original support, respectively. This numerical example illustrates the correct interpretation of the asymptotic results for fees, which is similar to the interpretation for distributions: linear fees are a good approximation not only in the limit, but also far away from the

[30]Sufficiency of Generalized mirror Pareto distributions for the optimality of linear fees follows from the argument in the text. Necessity was shown by Loertscher and Niedermayer (2007), a working paper superseded by the present paper. Building on this work, Niazadeh et al. (2014) extend the analysis of take-it-or-leave-it offers to linear fees that are close to optimal.

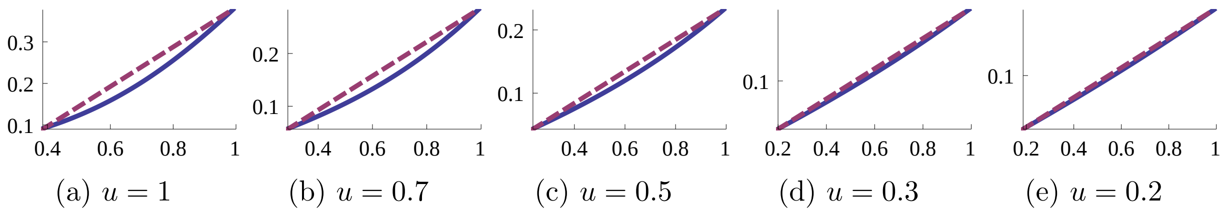


Figure 2: Optimal fee $\omega(\cdot)$ for truncated, rescaled distributions $F_u(v) = 1 - [1 - F_0(\overline{v} - u(\overline{v} - v))]/[1 - F_0(\overline{v} - u(\overline{v} - \underline{v}))]$, and $G_u(c) = G_0(\underline{c} + u(c - \underline{c}))/G_0(\underline{c} + u(\overline{c} - \underline{c}))$ for the same setup as in Figure 1; that is, for $u \in \{1, 0.7, 0.5, 0.3, 0.2\}$ for Beta distributions with support $[0, 1]$ and density $f(x) = g(x) \propto x^4(1 - x)^4$ (solid line) compared to an approximating linear fee (dashed).

limit.

To analyze and interpret the true – that is, denormalized – fee $\omega(p)$ that is approximately optimal away from but sufficiently close to the limit, it is also insightful to consider a j such that u_j^S and u_j^B are "sufficiently small" and to work with the denormalized limiting Pareto distributions

$$G^*(c) := \left(\frac{c - \underline{c}}{\overline{c} - \underline{c}}\right)^\sigma \quad \text{and} \quad 1 - F^*(v) := \left(\frac{\overline{v} - v}{\overline{v} - \underline{v}}\right)^\beta,$$

where $\underline{c} = \underline{v} = \underline{c}_j$ and $\overline{v} = \overline{c} = \overline{v}_j$ for notational simplicity. Observe that $G^*(c)$ has the linear virtual type function $\Gamma_\alpha^*(c) = c(1 + \alpha/\sigma) - \alpha\underline{c}/\sigma$, which implies that the fee is

$$\omega(p) = p - \Gamma_\alpha^{*-1}(p), \tag{8}$$

which is linear in p.

The expression in (8) offers a neat interpretation and simple comparative statics: A "Ramsey monopsonist" with valuation v who faces the supply $G^*(\hat{p})$ would set the price $\Gamma_\alpha^{*-1}(v)$.[31] A more elastic supply means a higher monopsony price and hence a lower fee. As an illustration, $\tilde{G}(\tilde{c}) = \tilde{c}^\sigma$ can be seen as isoelastic supply with elasticity σ. As the elasticity σ increases, the fee $\tilde{\omega}(\tilde{p}) = \alpha\tilde{p}/(\sigma + \alpha)$ decreases. Surprisingly, and interestingly, the fee is independent of distribution of valuations, which is a point we will return to below.

[31] A Ramsey monopsonist cares about his valuation v and a weighted average of the price p he pays and the seller's expected utility $E[c|c \leq p]$ and therefore maximizes $(v - (\alpha p + (1 - \alpha)E[c|c \leq p]))G^*(p)$, which implies the first-order condition $(v - \Gamma_\alpha^*(p))g^*(p) = 0$.

A final remark is in order before we proceed to the next section. One may be tempted to think that the linearity of the optimal fees in the limit stems from the fact that the range of the relevant, overlapping interval $[\underline{c}, \overline{v}]$ goes to zero with the standard reasoning that well-behaved functions are close to linear over short intervals. However, this is not so. The conditional distribution $\tilde{G}_j(\tilde{c})$ does, in general, not converge to a linear function (see Figure 1) nor does the optimal price function $\tilde{\Phi}^{-1}(\Gamma_\alpha(c))$ when $\delta > 0$. As shown in (8), the linearity of the optimal fee derives from the linearity of $\Gamma_\alpha(c)$, which in turn stems from the convergence of the distribution to (mirrored) Generalized Pareto distributions.

5 Discussion

In this section, we provide microfoundations for the transactions costs, comparative statics and extensions.

5.1 Microfoundations for the Transaction Costs

We now return to the microfoundation of transaction costs. There are various ways transaction costs may occur. Some types of transaction costs are simply exogenously given, such as transportation and moving costs. For additive transportation and moving costs K_j^B and K_j^S, the asymptotic results hold if $K_j^B + K_j^S \to \overline{v}_0 - \underline{c}_0$, that is, if the total transportation costs are close to the maximum possible gains from trade.

Other types of transaction costs can be viewed as the endogenous outcome of a larger model, as we describe in our empirical companion paper. No matter what the exact source of transaction costs, such transaction costs provide an explanation for why, for example, at any given point of time only a small percentage of real-estate properties are on the market or why Amazon considers the "tail of the distribution" (i.e. goods in thin markets, that are seldom traded) as one of its main sources of revenue.

Competing Direct Market Consider a competing direct market in the sense that the good can be traded at some price p, but this market is inefficient due to frictions: the probability of finding a trading partner is less than one and traders may have to spend potentially significant amount of time finding a trading partner. The intermediary's service is to enable trade with less frictions. So traders face the trade-off between either going to the intermediary and paying fees or going to the direct market and bare the costs of frictions. Examples would be buyers and sellers having the option to trade through other channels than eBay or Amazon. Another example is the real-estate market, where buyers and sellers have the option to rent rather than sell/buy, in which case p should be viewed as the net present value of rents.

Formally, consider a competing market that opens at regular intervals of length μ, with trade occurring at a price p. In the spirit of Satterthwaite and Shneyerov (2008), the rematching frequency μ can be viewed as a measure of efficiency of this market. The case $\mu \to \infty$ should be viewed as infinitely large frictions in the direct market, which is equivalent to the direct market not existing as an outside option, while $\mu \to 0$ means that frictions in the direct market vanish. Let the discount factor between two reopenings of the competing market (or, equivalently, the probability of not dropping out), be $\tilde{\delta} = e^{-\bar{\delta}\mu}$ where the parameter $\bar{\delta}$ is the discount rate, using Satterthwaite and Shneyerov (2007)'s terminology. Define a buyer's probability of trade at an instance when the competing market opens as $\tilde{\beta} = e^{-\bar{\beta}\mu}$ where $\bar{\beta}$ is buyers' trading rate. Analogously, define a seller's probability of trade $\tilde{\gamma} = e^{-\bar{\gamma}\mu}$ where $\bar{\gamma}$ is sellers' trading rate. The option value of trading in the competing market for a buyer with primitive valuation v_0 is $\sum_{t=0}^{\infty}(v_0 - p)\tilde{\beta}(1-\tilde{\beta})^t\tilde{\delta}^t = (v_0 - p)\beta$, where $\beta = \tilde{\beta}/[1-(1-\tilde{\beta}\tilde{\delta})]$ is the ultimate probability of trade in the competing market. Taking into account the option value of trading through the competing market, a buyer's effective valuation is $v = v_0 - \beta(v_0 - p) = \beta p + (1-\beta)v_0$, so that $K^B = \beta p$ and $\hat{K}^B = 1-\beta$. The upper bound of the buyer's effective valuation $\overline{v} = \beta p + (1-\beta)\overline{v}_0$ converges to p as μ converges to 0, that is, as the competing market becomes increasingly efficient. By a similar logic, the seller's effective cost is $c = c_0 + \gamma(p - c_0) = \gamma p + (1-\gamma)c_0$, where $\gamma = \tilde{\gamma}/[1-(1-\tilde{\gamma})\tilde{\delta}]$ is the seller's ultimate discounted probability of trade in the competing market. The lower bound of the seller's effective cost $\underline{c} = \gamma p + (1-\gamma)\underline{c}_0$ goes to p as $\mu \to 0$. Putting the pieces together, as the competing market becomes more efficient ($\mu \to 0$), the overlap of the seller's and the buyers' support $[\underline{c}, \overline{v}]$ shrinks towards p, which is what we need for our asymptotic results. Therefore, as frictions in the competing bilateral exchange decrease, an intermediary (whose service is to offer trade with less frictions) is forced to offer fees that are closer to linearity.

Dynamic Random Matching Model In most of our analysis, we are assuming that the buyer's valuation v and its distribution F are exogenously given. However, one can view our model as being embedded in a larger model, in which v and F are endogenous. Consider a dynamic random matching model in the spirit of Satterthwaite and Shneyerov (2008), which is here extended to include intermediaries: mass 1 of seller-intermediary pairs and mass 1 of buyers enter a market in every period. Buyers randomly choose a seller-intermediary pair to visit, which results in a Poisson distribution of the number of buyers a seller meets in every period. Sellers run an English auction with a reserve. If at least one bid is above the reserve, the seller-intermediary pair and the trading buyer leave the market. All participants who do not trade stay in the market. A steady-state equilibrium is one in which the mass and distribution of entering buyers equals the mass and distribution of exiting buyers, the same holding for sellers. The buyer's effective valuation is now endogenously given by $v = v_0 - \delta V_B(v_0)$, where $V_B(v_0)$ is the option value of future trade. One has to be additionally careful about the fact that buyers that trade with a lower probability stay longer in the market and are hence overrepresented in the pool of buyers in the market compared to entering buyers. While this endogeneity renders the model analytically intractable, in a previous version of this paper (Loertscher and Niedermayer, 2012), we solved this model numerically (under somewhat different assumptions).[32] We found that the linear fee approaches the optimal fee as the market becomes more efficient (which is modeled as δ increasing). The fundamental driving force is that an increase of δ reduces the upper bound of the buyer's effective valuation $\overline{v} = \overline{v}_0 - \delta V_B(\overline{v}_0)$ and hence reduces the overlap of supports $[\underline{c}, \overline{v}]$, which in turn makes linear fees closer to optimal. It turns out that the effect identified in the previous version of our paper is more general than the opportunity cost of future trade in dynamic random matching models: any transaction cost reducing gains from trade for the current

[32] We assumed that seller-intermediary pair write myopic contracts.

transaction makes optimal fees “more linear”.

5.2 Comparative Statics and Intuition

Our asymptotic results are interesting on their own, since they provide an explanation for the prevalent use of linear fees and also additional empirically testable implications, which we will discuss in detail in Subsection 5.4. But our asymptotic results also have a number of additional benefits. First, we discuss the benefits for comparative statics, which become much simpler in the limit. Using the results in the limit as a starting point, we can then derive comparative statics results away from the limit.

Comparative statics are most easily performed, and intuition for these developed, if one assumes a static setup with one buyer, that is, $\pi_1 = 1$ and $\delta = 0$. Replacing the primitive distributions and (virtual) types in (7) by F and G and v and c (and Φ and Γ_α^{-1}), the optimal fee with $\delta = 0$ is

$$\omega(p) = p - E[\Gamma_\alpha^{-1}(\Phi(v))|v \geq p].$$

Interpreting $G(p)$ and $1 - F(p)$ as quantities supplied and demanded at price p (see footnote 8), one can define the price elasticity of demand at v as $\eta_d(v) := vf(v)/(1-F(v))$ and the price elasticity of supply at c as $\eta_s(c) := cg(c)/G(c)$.

The comparative statics with respect to η_s and α are as expected: a global increase in $\eta_s(c)$ leads to lower fees and an increase in α leads to higher fees because $\Gamma_\alpha(c) = c(1 + \alpha/\eta_s(c))$ increases in α and decreases in $\eta_s(c)$.

The effect of the elasticity of demand η_d (and equivalently of $\Phi(v) = v(1-1/\eta_d(v))$) is clearly more complicated. It is useful to first consider the limit, in which G^* is Generalized Pareto, so that both Γ_α^* and the optimal fee ω are linear. Equation (8) shows that in this limit, the elasticity of demand η_d does not play any role for the fee whatsoever. This suggests that the way Γ differs from a linear function determines how η_d affects the optimal fee. As shown below, this is indeed the case and there is also an intuitive economic interpretation of the effects of the elasticity of demand.

Proposition 4. *Using the Taylor expansion of $\Gamma_\alpha^{-1}(x)$ around $\overline{v}$, the net price received* *by the seller is*

$$p - \omega(p) = \overbrace{\Gamma_\alpha^{-1}(p)}^{\textit{Ramsey monopsony price}} + \overbrace{\frac{[\Gamma_\alpha^{-1}(\overline{v})]''}{2} Var_{v\sim F}[\Phi(v) - \overline{v}|v \geq p]}^{\textit{second-order price endogeneity effect}}$$
$$+ \underbrace{\sum_{n=3}^{\infty} \frac{[\Gamma_\alpha(\overline{v})]^{(n)}}{n!} \{E_{v\sim F}[(\Phi(v) - \overline{v})^n|v \geq p] - E_{v\sim F}[\Phi(v) - \overline{v}|v \geq p]^n\}}_{\textit{higher-order price endogeneity effects}}, \quad (9)$$

where $[\Gamma_\alpha^{-1}(v)]^{(n)}$ *denotes the* n*-th derivative of* $\Gamma_\alpha^{-1}(v)$*.*

Recall that a Ramsey monopsonist with value x would set the price $\Gamma_\alpha^{-1}(x)$. Naturally, the intermediary's valuation for the good is p, so that absent any other effects the seller's net price should be $\Gamma_\alpha^{-1}(p)$, which is exactly what the seller receives when Γ_α is linear. However, for Γ_α nonlinear the price p is determined endogenously, which requires the optimal net price the seller receives to be adjusted.

According to Proposition 4, η_d has no first-order effect on the optimal fee. Indeed, as seen above when G is a mirrored Generalized Pareto distribution (which is equivalent to a linear Γ_α), $\omega(p)$ is independent of F. However, the second- and higher-order effects can go either way. To see this, assume $\alpha = 1$ and that $\Gamma_1^{-1}(x)$ is quadratic with a curvature $[\Gamma_1^{-1}(x)]'' = \overline{\gamma}_2$.[33] A quadratic form shuts down the higher-order effects and allows us to focus on the second-order price endogeneity effect. The fee is $\omega(p) = p - \Gamma^{-1}(p) - (\overline{\gamma}_2/2)\text{Var}[\Phi(v) - \overline{v}|v \geq p]$. For Γ_1^{-1} concave ($\overline{\gamma}_2 < 0$), an overall increase of the elasticity of demand can be shown to lead to an overall increase of the fee.[34]

These results are surprising and counterintuitive at first as one would expect more elastic demand to lead to lower fees. However, the intuition is that a more elastic demand causes the seller to lower the price excessively from the intermediary's point of view. To compensate for this, the optimal fee increases. If Γ_1^{-1} is convex, the converse occurs.

[33]It can be checked that a distribution G exists that generates a quadratic Γ_1^{-1} by inverting Γ^{-1} to get Γ and then solving the differential equation $\Gamma(c) = c + G(c)/g(c)$ with initial condition $G(\overline{c}) = 1$ for G. The closed-form solution for G is somewhat lengthy and hence not reported here.

[34]To see this, take distributions $\hat{F}$ and F with elasticities $\hat{\eta}_d(v)$ and $\eta_d(v)$ satisfying $\hat{\eta}_d(v) > \eta_d(v)$ for all $v < \overline{v}$. Because $\Phi(v) = v(1 - 1/\eta_d(v))$, this implies $\hat{\Phi}(v) > \Phi(v)$, which in turn implies $(\hat{\Phi}(v) - \overline{v})^2 < (\Phi(v) - \overline{v})^2$ for all $v < \overline{v}$. Further, F hazard rate dominates $\hat{F}$ because $\hat{f}(v)/(1 - \hat{F}(v)) = \hat{\eta}_d(v)/v > \eta_d(v)/v = f(v)/(1 - F(v))$. This implies $E[\hat{v}|\hat{v} \geq p] \leq E[v|v \geq p]$ for all p. Together with $(\hat{\Phi}(v) - \overline{v})^2 < (\Phi(v) - \overline{v})^2$, this implies $E[(\hat{\Phi}(\hat{v}) - \overline{v})^2|\hat{v} \geq p] \leq E[(\Phi(v) - \overline{v})^2|v \geq p]$. Therefore, fees are higher with $\hat{F}$ than with F, since $\gamma_2 < 0$ and $\text{Var}[\Phi(v) - \overline{v}|v \geq p] = E[(\Phi(v) - \overline{v})^2|v \geq p] - E[\Phi(v) - \overline{v}|v \geq p]^2 = E[(\Phi(v) - \overline{v})^2|v \geq p] - (p - \overline{v})^2$.

Whether a concave or a convex Γ_1^{-1} is more common case in real world markets is an open question. The only empirical paper that estimates the virtual cost function that we are aware of is Loertscher and Niedermayer (2017), which uses data from the real estate brokerage market in Boston in the early 1990s. The estimated virtual cost function is very close to linear, but slightly convex. This implies a slightly concave inverse virtual cost function Γ_1^{-1}.

Important contributions by Bulow and Pfleiderer (1983), Aguirre et al. (2010), Bulow and Klemperer (2012), and Weyl and Fabinger (2013) have identified a number of properties of the demand function such as its curvature, the curvature of the inverse demand function, the pass-through rate, and the markup- or quantity-weighted average pass-through, which prove useful in a variety of contexts in industrial organization. Naturally, one may then wonder whether the counterintuitive result that optimal fees are sometimes higher for a higher elasticity of demand may be explained by alternative properties of the demand function. The example with Γ_1^{-1} quadratic shows that the answer is no. Any change of any property of the demand function F that leads to higher fees for $\overline{\gamma}_2 < 0$ will lead to lower fees for $\overline{\gamma}_2 > 0$ and will have no effect on the fees for $\overline{\gamma}_2 = 0$.

Our results are also relevant for public finance in the context of indirect taxation.[35] In thick markets, it is well known that less elastically demanded goods should be taxed more heavily (see Salanié, 2003). For thin markets, our results imply that the elasticity of supply is key. Depending on the curvature of Γ_α^{-1}, one should either tax the good whose demand is more elastic or the one whose demand is less elastic. Our results are also of relevance for the practice of competition policy. When there is suspicion that fee-setting intermediaries, such as auction houses and platforms or real-estate brokers, collude, the standard approach would be to estimate the demand function and to then evaluate how closely prices are to the monopoly price implied by the estimates. Our analysis suggests that in thin markets with intermediaries, the first look should be at the

[35]Indirect taxes are often different for different product categories. As an example, the EU financial transaction tax levies 0.1% on share transactions and 0.01% on transactions involving derivatives. Value added taxes and sales taxes in many countries differ across products, with some goods being exempt from indirect taxes altogether.

elasticity of *supply* rather than the elasticity of demand. As a first-order approximation, the elasticity of demand does not matter for the fees of a profit maximizing intermediary. Instead, colluding intermediaries should be expected to leave a net price to the seller which corresponds to the price set by a monopsonist whose valuation is the gross price (again, as a first-order approximation).

5.3 Extensions

The setup we study is amenable to a variety of interesting and natural extensions. The limiting Pareto distribution turns out to have interesting implications in these extensions. Due to space constraints, we will only sketch what we consider to be the most valuable ones.

Non-Stationarity Let us first briefly explain how our analysis can be extended to account for non-stationary environments at what is essentially a cost in notation. Assuming that the sequences of time varying discount factors δ_t, distributions F_t and random arrival processes π_B^t are known, one can proceed in analogy to the way we proceeded under the assumption of stationarity. Although the optimal transaction fee ω_t in period t will in general vary over time because of non-stationarity, a simple argument based on what we call "expectational fees" (which are defined and derived in Lemma 2 in the Appendix) shows that, in the limit as markets become thinner, the optimal (normalized) transaction fees will be linear and stationary in the limit, too. The limit results also hold in a setup in which the distribution of buyers' valuations changes stochastically over time. Appendix B contains more details. The stationarity of the optimal limiting fees is particularly remarkable because the optimal reserve price path chosen by the seller will in general be non-stationary.

Linear Fees, First-Price Auctions, and the Informed Principal Problem Given linear transaction fees ω, the payoff of the seller upon selling at some price $\breve{p}$ is linear in the transaction price. Consequently, linear fees correspond to the case where the seller is a risk-neutral agent with a linear Von-Neuman-Morgenstern utility function. As is well known, with risk-neutral agents the revenue equivalence theorem applies.[36] This implies that, given linear fees, using a first-price auction in which the seller sets the reserve is equivalent to the second-price auction we have assumed thus far. Moreover, due to the results for the informed principal problem in linear environments with independent private values of Mylovanov and Tröger (2014), keeping fixed the linear fee the expected payoffs conditional on type would be unchanged if the seller could choose the trading mechanism after having learned his type (in any strongly neologism-proof perfect Bayesian equilibrium). Thus, with linear fees the broker could even delegate the choice of the mechanism to the seller. This has the important implication that the intermediary could raise his revenues in a rather decentralized way: he simply sets the linear fees and can leave the choice of the details of the bargaining protocol to the seller.

5.4 Empirical Implications

Our theoretical results have empirical implications that are consistent with observations in several markets.

One implication is that in a static model with an English auction (which is a reasonable approximation of auction houses, such as Sotheby's and Christie's, and online auction sites, such as eBay), the fee is independent of the number of bidders in the auction. This result is similar in spirit to the well-known result that the profit maximizing reserve price in an English auction with private values is independent of the number of bidders. More importantly, it is consistent with the observation that the fees of auction houses and platforms do not depend on the number of bidders.

Further, if the optimal fees are linear, which is rationalized in our model with the distribution of sellers' costs G being well approximated by a Generalized Pareto distribution, then the fee should be independent of the number of buyers and the distribution of their valuations even in a dynamic model. Empirically, one observes that indeed the fees of real-estate agents and of auction houses and platforms exhibit little variation over time and across regions (see Hsieh and Moretti (2003) who report the surprising invari-

[36] The only additional assumption with a random number of bidders is that bidders be symmetrically informed about the number of other bidders participating (see e.g. Krishna, 2002, chapter 3).

ance of real estate brokerage fees in the US). This prediction does not even require the distribution G to be invariant: it is sufficient that the lower bound of support $\underline{c}$ remains close to zero and that the Pareto tail index σ does not change much.[37]

It is also remarkable that Amazon charges linear fees for 33 out of 38 categories of goods that are sold by third-parties through Amazon's website. If linear fees were simply due to simplicity, as is sometimes argued, one would not expect such a large number of categories.

For the specific example of real estate brokerage fees, a number of additional observations are consistent with the predictions of our theory. First, sellers with a higher loan-to-value ratio ask higher prices (Genesove and Mayer, 1997). Second, sellers who had bought their houses when average real estate prices were high, ask for higher prices than those who had bought when prices were low (Genesove and Mayer, 2001). Third, quality-adjusted prices and time on market of houses are positively correlated in cross-sectional data and negatively correlated in longitudinal data. Fourth, direct sellers – that is, private owners who sell their houses without a broker or brokers who sell their own houses – sell at higher prices than private owners who sell through a broker (see Rutherford et al., 2005; Levitt and Syverson, 2008; Hendel et al., 2009). All these observations are either predicted by or at least consistent with our theory.[38]

In our empirical companion paper, we go beyond explaining known stylized facts with our theory: we compare the novel predictions of our theory with the data. Among other things, we consider the surprising theoretical prediction that sellers' costs are approximately Generalized Pareto distributed. We find that Generalized Pareto distributions are indeed a reasonable approximation of sellers' costs in the sense that the 6 per cent

[37] It is still an open question why the Pareto tail index exhibits little variation. But this is a puzzle that has been observed in a variety of contexts, such as the relative inertia of the Pareto tail index of income and wealth distributions or of the distribution of city sizes.

[38] The first and the second observation can be explained by the seller's opportunity cost of selling being correlated with the loan-to-value ratio and the seller's valuation at the previous transaction, respectively. For the third observation, the cross-sectional prediction is simply due to higher prices leading to a longer time on market. For the longitudinal prediction, observe that in a static model, the seller's reserve price is independent of the number of bidders arriving, which naturally generates high prices and short times on market when there are many buyers. It generates the opposite if there are few buyers. We show in Proposition 4 of our companion paper (Loertscher and Niedermayer (2017)) that the fourth stylized fact is consistent with our theory.

fee achieves more than 98% of what the optimal non-linear fee could achieve.

5.5 Alternative Modelling Approaches

We now briefly discuss a number of alternative modelling approaches and the advantages of the approach we took relative to those.

Why not a Static Model? Some of the economic insights from our model can also be obtained in a static setup ($\delta = 0$). Hence, one might wonder whether a static model would have been sufficient. We think that a dynamic model is important for several reasons. First, as discussed, a straight-forward generalization of our results to a non-stationary setup shows that even in a non-stationary setup the asymptotic optimal fee is, quite surprisingly, stationary. Second, the fact that sellers offer their good for sale in multiple periods is an important feature of many real world markets. Third, some of the microfoundations for the transaction costs provided above are most naturally expressed in a dynamic environment. Last, the empirical analysis of real-estate brokerage fees in our companion paper uses data on time on market, which is only meaningful in a dynamic setup. The time on market is a crucial ingredient in the empirical identification strategy in our companion paper.

Why not Linear Pricing? Linear fees in intermediated markets may appear similar to the more familiar concept of linear pricing in the usual (non-intermediated) markets. However, the similarity is superficial and misleading. First, linear fees are used in settings with single-unit supplies and demands whereas linear pricing refers to prices that are linear in quantity. Second, the linearity of the fee refers to linearity in *price* rather than quantity. Both are crucial differences.[39] Further, the explanations often provided for linear pricing, such as the possibility of arbitrage, competition precluding price dis-

[39] The difference between linear pricing in quantity and linear fees in prices is fundamental. Production costs may plausibly be proportional to quantity, whereas costs of intermediation are unlikely to be proportional to the price. Further, we use the term "linear fee" to describe a fee that is possibly a fixed fee plus a percentage of the price just like it is customary to refer to demand functions with the same form as representing linear demand whereas in fact these are affine functions. In contrast, "linear pricing" refers to pricing that is *proportional* to the quantity.

crimination, and simplicity are fundamentally different from what drives fees to linearity in our setting.[40]

Why Not One-Sided Private Information? One might ask whether a simpler model with one-sided private information would not do. We now argue that the answer is negative. For this purpose, consider the following static model with one-sided private information model: the seller's cost c is public information, there is only one buyer whose valuation v is private information and drawn from the distribution F, and the intermediary maximizes profits ($\alpha = 1$).[41] We will discuss how results carry over to a setting with multiple buyers later on.

Since the seller has no private information, the intermediary need not give him any rents and will make sure that the seller gets c in expectations in case of trade. We know from Myerson (1981) that the optimal mechanism to sell to a buyer is to set the price $\Phi^{-1}(c)$. This implies a simple implementation of the profit maximizing mechanism: the mechanism requires the seller to set the price $\Phi^{-1}(c)$ and charges a transaction fee $\omega(p)$ as a function of the transaction price p that makes sure that the seller gets the net price c (the transaction price always happens to be $p = \Phi^{-1}(c)$). Any arbitrary fee function works, as long as $\omega(\Phi^{-1}(c)) = \Phi^{-1}(c) - c$. Take e.g. linear fees $\hat{\omega}(p) = \mu(\Phi^{-1}(c) - c) + (1 - \mu)(1 - c/\Phi^{-1}(c))p$ for some $\mu \in [0, 1]$. $\hat{\omega}$ is constant for $\mu = 0$ and a percentage fee for $\mu = 1$.

[40] One explanation for linear pricing is the possibility of arbitrage: if two units of a good are less than twice as expensive as one unit, a buyer might buy two units, consume one unit and resell the other. However, it should be clear that this does not apply to intermediaries: if the percentage fee were different for a house that sells at \$200,000 than for a house that sells at \$400,000, there is no possibility of arbitrage for the seller. Another explanation given for linear pricing is that sufficient competition makes price discrimination impossible. However, as shown in the preceding analysis, *linear fees* in contrast to *linear pricing* do not mean the absence of price discrimination. A further statement sometimes made about linear pricing is that there is not really any economic explanation for linear prices, they are just plain simpler. Leaving aside that this "theory" is open to the Lucas critique, for an intermediary's fees it is also not a particularly convincing explanation. As mentioned, Amazon has an elaborate pricing scheme in which it charges different fees for 38 different categories of goods. 33 out of 38 categories have linear fees (i.e. a fixed fee plus a percentage of the transaction price).

[41] The setup with two-sided private information provides a coherent, internally consistent framework to analyze indirect taxation. Without private information about the seller's cost (and no fixed costs) and without imposing exogenously given constraints in policy instruments, the government could achieve first-best by forcing sellers to price at marginal costs. However, such large scale intervention across many different industries seems a daunting task for any government.

This simple model illustrates why a one-sided private information model would not have been sufficient. First of all, it has the implication that a profit maximizing intermediary would leave zero utility to the seller.[42] In its basic form, it would also not explain the prevalence of price dispersion in many markets. Such a model would also not lend itself to an analysis about the structure of fees since the fee structure is irrelevant (e.g. the value of μ in $\hat{\omega}$ does not matter).

This also illustrates why a one-sided private information model – in contrast to our model – gives the prediction that the fee decreases with the elasticity of demand $\eta_d(v) = vf(v)/(1 - F(v))$: it is easy to check that the linear fee $\hat{\omega}$ provided as an example decreases with η_d because the monopoly price $\Phi^{-1}(c)$ decreases with elasticity.[43]

With multiple buyers, the same reasoning holds as with one buyer, except that now the *expected* fee has to be equal to the *expected* profit from selling minus the seller's cost c. Results also carry over to multiple periods.

The result that with one-sided private information the structure of the fee is not pinned down extends to thick markets with two-sided private information because in a thick market there is no uncertainty about the realized distribution of sellers' costs. Consequently, in thick markets the cost of the marginal seller is public information (see Niedermayer and Shneyerov (2014)).

5.6 Services of Intermediaries

We have made a conscious modeling choice in this paper by modelling optimal pricing behavior by intermediaries rather than the specifics of the services of intermediaries. There are good reasons for this choice. There is a wide variety of intermediaries charging fees, such as real-estate brokers, Amazon, eBay, and iTunes, offering a wide variety of services. If we add taxation by governments into the picture, there is even more

[42] Even if the seller's cost is not known to the intermediary, but known to other market participants, the same results hold because the intermediary can extract information about the seller's cost costlessly by a Crémer and McLean (1988) type of mechanism. Some of the one-sided private information papers restrict the set of mechanisms the intermediary can choose in such a way that in the restricted set fees other than percentage fees are suboptimal.

[43] Technically speaking, the implication is that the fee $\omega(\Phi^{-1}(c)) = \Phi^{-1}(c) - c$ paid at the equilibrium price decreases with η_d. Since the fee for other prices is arbitrary and irrelevant, one could also construct fees that decrease at the equilibrium price, but increase at some off-equilibrium prices.

heterogeneity. There are a variety of mutually non-exclusive explanations as to why the services of private intermediaries and governments are useful, such as for reducing transaction and search costs, certifying quality, improving matching, building reputation, providing infrastructure that facilitates trade, and enforcing contracts.[44] Rather than going through all the combinations of such explanations, which will be special for every industry (and even for a single industry, there is typically no consensus on which services are the most important), we provide a general model and focus on what is common to all intermediaries: that they raise revenues by charging transaction fees. This approach does come at a price: our theory remains silent on what is special about the services of one or another type of intermediary. However, we believe this price is worth paying, since we gain a deeper, more general understanding of the determinants of fees. There is precedent of such an approach paying off: industrial organization has developed a number of theoretical and structural estimation tools to deal with optimal pricing by one-sided (i.e. non-intermediary) firms. These general tools have turned out to be very useful, despite the fact that optimal pricing is done by very different firms offering very different products such as cereals, cars, and pharmaceutical products (see e.g. Berry et al. (1995) and the subsequent literature) and despite the fact that an optimal pricing approach remains silent about the question why consumers buy cereals, cars, and pharmaceutical products.

5.7 Principal Agent Theory of Intermediation

Another related question is whether the fees charged by real-estate brokers should alternatively be viewed as the solution of a principal-agent problem. Indeed, economists often casually refer to real-estate brokerage as an example of a principal-agent problem. The (informal) theory typically looks like the following. A seller hires a real-estate agent and incentivizes him to drive up the sales price by offering a fee that increases with the price. While this theory is superficially appealing, it stands in stark contradiction with many

[44]See Spulber (1999) and Salanié (2003) for an overview of the role of private intermediaries and governments, respectively.

empirical observations made in the real-estate brokerage market described above.[45] The many puzzles that appear when viewing real-estate brokerage fees from a principal-agent perspective all disappear if one is willing to take a novel approach and consider brokerage fees from an optimal pricing perspective. If one further has the ambition to have a theory that explains the fees not only of real-estate brokers, but also of Amazon, eBay, and possibly even the taxes levied by governments, then a principal-agent explanation becomes even less convincing: it is e.g. hard to believe that the reason why eBay's fees increase with the transaction price is that sellers are trying to incentivize eBay to exert effort to raise the price.

However, besides raising prices, the role of intermediaries may also be to enable transactions (or increase the probability of a transaction). Most of our analysis is orthogonal to the notion that fees serve to incentivize brokers to enable transactions, as we illustrate in the following with a simple principal-agent extension of our model. Assume that the broker can exert effort e, which affects the probability $\pi(e)$ that the seller meets suitable buyers, where $\pi' > 0$, $\pi'' < 0$. The broker exerts effort to maximize his expected profit $\pi(e)E_{c\sim G}[W_I(c, \boldsymbol{P}(c), \boldsymbol{\omega})] - e$, where $E_{c\sim G}[\cdot]$ is the broker's expected profit in case that suitable potential buyers are found. The broker will maximize his profits by choosing effort level $e^* = (\pi')^{-1}(E_{c\sim G}[W_I(\cdot)])$. Therefore, α can be seen as affecting a trade-off between two distortions. A higher α increases $E_{c\sim G}[W_I(\cdot)]$ and leads to more effort by the broker and hence a higher probability that the seller and some buyers meet, but decreases the probability that trade occurs conditional on the seller and buyers meeting. Our theory can then be interpreted as saying that the intermediary and the seller write a contract in which they choose α such that it induces the intermediary to set the optimal effort level e.

[45] To mention only one of the many observations, consider a transaction in which both the seller and the buyer have a broker. For such transactions, both brokers get 3% of the transaction price. This means that the buyer's broker gets *more* if the transaction price is *higher*, that is if the outcome is *less* favorable from the buyer's perspective, something that cannot be explained by a principal-agent theory. Note that this observation holds not only for co-operating agents on the buyer's side (who have to act in the buyer's interest), but also for buyers' agents (who are contractually obliged to act in the buyer's fiduciary interests). See Loertscher and Niedermayer (2017) for more details of why the principal-agent theory is unconvincing as an explanation of real-estate brokerage fees.

5.8 Equivalence Results

As discussed in Section 5.5, our two-sided private information setting avoids the multiplicity of implementations present in one-sided private information settings of fees and taxes, i.e. in our setting fixed fees, percentage fees, and non-linear fees are not equivalent. However, our setting still inherits another well-known equivalence result from the tax incidence literature: only the sum of the seller's and the buyer's fee matters, that is, it is irrelevant whether the seller, the buyer, or both pay (parts of) the transaction fees (or taxes). For example, it is equivalent to our setting if the net rather than the gross price is defined to be the transaction price and the buyer rather than the seller is said to pay the fee.

A further equivalence result is that our model can be seen both as a model of non-durable goods (which provide a one-off consumption utility when consumed) and of durable goods (which provide a consumption utility in each period they are consumed). A durable goods setting translates in a straightforward way into our setting with valuations (or opportunity costs of selling) being the net present value of consumption utilities.[46]

6 Conclusions

We provide a parsimonious theory of optimal transaction fees in thin markets and show that the optimal fees converge to linear fees as markets become increasingly thin. Our companion paper shows that, given empirically estimated distributions, little if anything is lost through the use of linear fees. This suggests that, going forward, researchers and other analysts of transaction fees in thin markets might as well assume (generalized) Pareto distributions, which are analytically convenient and equivalent to linear fees being optimal.

While the main purpose of this article is to develop a general model of transaction

[46]To see this, consider the non-durable good setup in our model, in which a seller of type c has a payoff of 0 if he never trades and an expected discounted payoff of $\delta^T(p-c)$ if he trades at price p in period T. Alternatively, we could assume that the good is durable, i.e., the good gives the seller a utility of $\tilde{c}$ in every period in which he consumes it. Under this assumption, if he never sells, his expected discounted utility is $\tilde{c}/(1-\delta)$. In contrast, if he sells at price p in period T, his expected discounted payoff is $\delta^T p + \tilde{c}(1-\delta^T)/(1-\delta)$. Taking the difference, one obtains $\delta^T(p-\tilde{c}/(1-\delta))$. Thus, for $c = \tilde{c}/(1-\delta)$ the two models are isomorphic.

fees as optimal pricing, a positive side effect of having such a theory is that it resolves many of the puzzling observations documented in the empirical literature on real-estate brokerage fees by providing an alternative to the principal-agent view.

A further aspect of our model that deserves emphasis is that in thin markets optimal fees vary little with the underlying environment. In real-estate brokerage, the invariance of the 6 per cent fees across times and markets is a well-documented stylized fact (see e.g. Hsieh and Moretti, 2003). According to our theory, the asymptotically optimal fee in thin markets is linear and independent of demand-side factors. The asymptotic optimal fee depends on the Pareto tail index σ. The invariance of fees is hence ultimately related to the invariance of the Pareto tail index. This invariance has been documented in a number of empirical settings, such as for income and wealth distributions, the sizes of cities, and the strengths of earthquakes.[47]

References

Aguirre, I., S. Cowan, and J. Vickers (2010): "Monopoly Price Discrimination and Demand Curvature." *American Economic Review*, 100, 1601 – 1615.

Anderson, S., A. De Palma, and B. Kreider (2001a): "Tax Incidence in Differentiated Product Oligopoly," *Journal of Public Economics*, 81, 173–192.

——— (2001b): "The Efficiency of Indirect Taxes under Imperfect Competition," *Journal of Public Economics*, 81, 231–251.

Baliga, S. and R. Vohra (2003): "Market Research and Market Design," *Advances in Theoretical Economics*, 3, Article 5.

Balkema, A. and L. De Haan (1974): "Residual Life Time at Great Age," *Annals of Probability*, 2, 792–804.

Barvinek, E., I. Daler, and J. Francu (1991): "Convergence of Sequences of Inverse Functions," *Archivum Mathematicum*, 27, 201–204.

[47] See Gabaix (2016) for a general discussion of power laws in Economics.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 841–890.

Board, S. (2008): "Durable-Goods Monopoly with Varying Demand," *Review of Economic Studies*, 75, 391–413.

Brander, J. A. and A. Zhang (1990): "Market conduct in the airline industry: an empirical investigation," *The RAND Journal of Economics*, 567–583.

Bulow, J. and P. Klemperer (2012): "Regulated Prices, Rent Seeking, and Consumer Surplus." *Journal of Political Economy*, 120, 160 – 186.

Bulow, J. and J. Roberts (1989): "The Simple Economics of Optimal Auctions," *Journal of Political Economy*, 97, 1060–90.

Bulow, J. I. and P. Pfleiderer (1983): "A Note on the Effect of Cost Changes on Prices." *Journal of Political Economy*, 91, 182 – 185.

Caillaud, B. and B. Jullien (2003): "Chicken & Egg: Competition among Intermediation Service Providers," *RAND Journal of Economics*, 34, 309–328.

Corts, K. S. (1999): "Conduct parameters and the measurement of market power," *Journal of Econometrics*, 88, 227–250.

Crémer, J. and R. McLean (1988): "Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56, 1247 – 1257.

De Haan, L. and A. Ferreira (2006): *Extreme Value Theory: An Introduction*, Springer Verlag.

Delipalla, S. and M. Keen (1992): "The comparison between ad valorem and specific taxation under imperfect competition," *Journal of Public Economics*, 49, 351 – 367.

Department of Justice (2007): "Competition in the Real Estate Brokerage Industry," *Report by the Federal Trade Commission and the U.S. Department of Justice.*

Falk, M., J. Hüsler, and R. Reiss (2010): *Laws of Small Numbers: Extremes and Rare Events*, Springer Verlag.

Gabaix, X. (2016): "Power Laws in Economics: An Introduction," *Journal of Economic Perspectives*, 30, 185–206.

Garrett, D. F. (2016): "Intertemporal price discrimination: Dynamic arrivals and changing values," *American Economic Review*, 106, 3275–3299.

Genesove, D. and C. J. Mayer (1997): "Equity and Time to Sale in the Real Estate Market," *American Economic Review*, 87, 255–269.

——— (2001): "Loss Aversion and Seller Behavior: Evidence from the Housing Market," *Quarterly Journal of Economics*, 116, 1233–1260.

Gomes, R. (2014): "Optimal Auction Design in Two-Sided Markets," *RAND Journal of Economics*, 45, 248–272.

Gresik, T. (1991): "Ex Ante Efficient, Ex Post Individually Rational Trade," *Journal of Economic Theory*, 53, 131–145.

Hagiu, A. (2007): "Merchant or Two-Sided Platform?" *Review of Network Economics*, 6, 115–33.

Hendel, I., A. Nevo, and F. Ortalo-Magné (2009): "The Relative Performance of Real Estate Marketing Platforms: MLS versus FSBOMadison.com," *American Economic Review*, 99, 1878 – 98.

Hsieh, C.-T. and E. Moretti (2003): "Can Free Entry Be Inefficient? Fixed Commissions and Social Waste in the Real Estate Industry," *Journal of Political Economy*, 111, 1076–1122.

Johnson, J. P. (2014): "The agency model and MFN Clauses," *Available at SSRN 2217849*.

Jullien, B. and T. Mariotti (2006): "Auction and the Informed Seller Problem," *Games and Economic Behavior*, 225–258.

Krishna, V. (2002): *Auction Theory*, Elsevier Science, Academic Press.

Levitt, S. D. and C. Syverson (2008): "Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions," *Review of Economics and Statistics*, 90, 599–611.

Loertscher, S. and A. Niedermayer (2007): "When is Seller Price Setting with Linear Fees Optimal for Intermediaries?" *Working Paper, University of Bern.*

——— (2012): "Fee Setting Intermediaries: On Real Estate Agents, Stock Brokers, and Auction Houses," *Working Paper.*

——— (2017): "Percentage Fees in Thin Markets: An Empirical Perspective," Mimeo.

——— (2018): "Entry-Deterring Agency," Mimeo, University of Melbourne.

Matros, A. and A. Zapechelnyuk (2008): "Optimal Fees in Internet Auctions," *Review of Economic Design*, 12, 155–63.

Myerson, R. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6, 58–73.

Myerson, R. and M. Satterthwaite (1983): "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, 29, 265–281.

Mylovanov, T. and T. Tröger (2014): "Mechanism Design by an Informed Principle," *Working Paper.*

Niazadeh, R., Y. Yuan, and R. Kleinberg (2014): "Simple and Near-Optimal Mechanisms For Market Intermediation," Http://arxiv.org/abs/1409.2597v1.

Niedermayer, A. and A. Shneyerov (2014): "For-Profit Search Platforms," *International Economic Review*, 55, 765–789.

Pickands, J. (1975): "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, 119–131.

Riley, J. and R. Zeckhauser (1983): "Optimal Selling Strategies: When to Haggle, When to Hold Firm ," *Quarterly Journal of Economics*, 267–89.

Rutherford, R., T. Springer, and A. Yavas (2005): "Conflicts between principals and agents: evidence from residential brokerage," *Journal of Financial Economics*, 76, 627–665.

Salanié, B. (2003): *The Economics of Taxation*, MIT Press.

Satterthwaite, M. and A. Shneyerov (2007): "Dynamic Matching, Two-Sided Information, and Participation Costs: Existence and Convergence to Perfect Competition," *Econometrica*, 75, 155–200.

——— (2008): "Convergence to Perfect Competition of a Dynamic Matching and Bargaining Market with Two-sided Incomplete Information and Exogenous Exit Rate." *Games and Economic Behavior*, 63, 435–467.

Shy, O. and Z. Wang (2011): "Why Do Payment Card Networks Charge Proportional Fees?" *American Economic Review*, 101, 1575–1590.

Spulber, D. F. (1999): *Market Microstructure: Intermediaries and the Theory of the Firm*, Cambridge University Press, Cambridge.

Tirole, J. (2016): "From Bottom of the Barrel to Cream of the Crop: Sequential Screening With Positive Selection," *Econometrica*, 84, 1291–1343.

Wang, Z. and J. Wright (2017): "Ad-valorem platform fees and efficient price discrimination," *RAND Journal of Economics*, 48, 467–484.

Weyl, G. and M. Fabinger (2013): "Pass-Through as an Economic Tool: Principle of Incidence under Imperfect Competition," *Journal of Political Economy*, 121, 528 – 583.

Yavas, A. (1992): "Marketmakers versus Matchmakers," *Journal of Financial Intermediation*, 2, 33–58.

Appendix

A Proofs

A.1 Propositions 1, 3, and 4

Proof of Proposition 1. The first order condition for the seller's maximization problem is

$$[(R_\omega(p)-c)(1-F_\infty(p))]' = -[\tilde{\Phi}_\omega(p)-c]f_\infty(p)$$

with

$$\tilde{\Phi}_\omega(p) := R_\omega(p) - R'_\omega(p)\frac{1-F_\infty(p)}{f_\infty(p)}. \tag{10}$$

We will show that the expression for $\tilde{\Phi}_\omega$ in (10) is the same as the one in the proposition.

First, observe that

$$R_\omega(p) = \frac{(p-\omega(p))(F_{(2)}(p)-F_{(1)}(p)) + \int_p^{\overline{v}}(v-\omega(v))dF_{(2)}(v)}{1-F_{(1)}(p)}$$

can be rewritten as

$$R_\omega(p) = \frac{\int_p^{\overline{v}}\Phi_\omega(v)dF_{(1)}(v)}{1-F_{(1)}(p)}$$

where

$$\Phi_\omega(p) := p - \omega(p) - (1-\omega'(p))\frac{1-F(p)}{f(p)}$$

That the two expressions for R_ω are equal can be checked by observing that $R_\omega(\overline{v}) = \overline{v}$ for both expressions and that the derivatives $[R_\omega(p)(1-F_{(1)}(p))]'$ can be shown to be equal for both expression for R_ω with some algebra and by using the fact[48]

$$\frac{F_{(2)}(p)-F_{(1)}(p)}{f_{(1)}(p)} = \frac{1-F(p)}{f(p)}.$$

One can also show with some algebra that

$$R'_\omega(p) = \frac{f_{(1)}(p)}{1-F_{(1)}(p)}(R_\omega(p)-\Phi_\omega(p)) \tag{11}$$

and that

$$\frac{f_{(1)}(p)}{1-F_{(1)}(p)}\frac{1-F_\infty(p)}{f_\infty(p)} = \frac{1-\delta F_{(1)}(p)}{1-\delta} \tag{12}$$

[48] This is easily seen to be true once one notes that $f_{(1)}(v)$ can be written as $f_{(1)}(v) = f(v)\sum_{B=1}^{\infty}\pi_B BF(v)^{B-1}$ and by noticing that $F_{(2)}(v) - F_{(1)}(v) = (1-F(v))\sum_{B=1}^{\infty}\pi_B BF(v)^{B-1}$.

Plugging (11) and (12) into (10) yields

$$\tilde{\Phi}_\omega(p) = R_\omega(p) - (R_\omega(p) - \Phi_\omega(p))\frac{1-\delta F_{(1)}(p)}{1-\delta}$$

the derivative of which can be rearranged to

$$\tilde{\Phi}'_\omega(p) = \frac{1-\delta F_{(1)}(p)}{1-\delta}\Phi'_\omega(p)$$

Since this expression for $\tilde{\Phi}'_\omega$ is equal to the expression for $\tilde{\Phi}'_\omega$ in the proposition and since the two expressions for $\tilde{\Phi}_\omega(p)$ are equal to $\overline{v}$ for $p = \overline{v}$, Φ_ω is the same in the proposition as in (10).

Therefore, the seller's first-order condition can be written as $-(\tilde{\Phi}_\omega(p) - c)f_\infty(p) = 0$, which implies the optimal price $\tilde{\Phi}_\omega^{-1}(c)$ for a seller with cost c as stated in the proposition.

□

Proof of Proposition 3. The proof of part (i) relies on Extreme Value Theory, which we summarize in Appendix C. Φ continuously differentiable implies that, for some constant $\overline{\beta}$,

$$\lim_{v\to\overline{v}} \frac{d}{dv}\left[\frac{1-F(v)}{f(v)}\right] = \lim_{v\to\overline{v}} \frac{d}{dv}\left[v - \Phi(v)\right] = \overline{\beta}. \tag{13}$$

Equation (13) is the von Mises condition as stated in Theorem 2 in Appendix C. By Theorem 2, this implies that F is in the domain of attraction of an extreme value distribution (see Definition 1). By the Pickands-Balkema-de Haan Theorem (see Theorem 1), this in turn implies that F has a Generalized Pareto upper tail as defined in (24). This implies uniform convergence of the normalized distribution $\tilde{F}_j$ to $\tilde{F}^*(\tilde{v}) = 1-(1-\tilde{v})^\beta$, because of the definition of the normalized variable $\tilde{v}$. Analogous reasoning applies for the convergence of $\tilde{G}_j$.

Proof of part (ii): First, define $\bar{\beta} := \lim_{v\to\overline{v}} 1 - \Phi'(v)$, $\bar{\sigma} := \lim_{c\to\underline{c}} \Gamma'(c) - 1$, $\beta := -1/\overline{\beta}$, and $\sigma := 1/\overline{\sigma}$. Observe that by l'Hôpital's rule

$$\lim_{v\to\overline{v}} \frac{(\overline{v}-v)f(v)}{1-F(v)} = \lim_{v\to\overline{v}} \frac{\overline{v}-v}{v-\Phi(v)} = \lim_{v\to\overline{v}} \frac{-1}{1-\Phi'(v)} = \beta.$$

The following two constructs are used in the remainder of the proof: First, instead of setting a reserve price p that leads to expected revenue $k = R(p)$ conditional on trade in this period, one can alternatively and hypothetically assume that the seller sets an expected transaction price k, conditional on trade ever occurring, that leads to trade with probability $1 - \overline{F}(k) := 1 - F_\infty(R^{-1}(k))$. Second, the intermediary can levy an "expectational fee" $\varpi(k)$ on the expected transaction price k. The following lemma derives the expectational fee $\varpi(k)$ that implements the allocation rule derived in Lemma 1.

Lemma 2. *The expectational transaction fees that implement the optimal mechanism described in Lemma 1 are*

$$\varpi(k) = k - \frac{\int_k^{\overline{v}} \Gamma_\alpha^{-1}(\overline{\Phi}(v))\overline{f}(v)dv}{1 - \overline{F}(k)}.$$

Proof of Lemma 2. The expected profit of a seller with cost who faces a fee ϖ is

$$(1 - \overline{F}(k))(k - \varpi(k) - c).$$

Substituting $\varpi(k)$ by the expression in Proposition 2, the maximization problem becomes

$$\max_k \int_k^{\overline{v}} \Gamma_\alpha^{-1}(\overline{\Phi}(v))\overline{f}(v)dv - (1 - \overline{F}(k))c.$$

The first-order condition is

$$0 = -\overline{f}(k(c)) \left[\Gamma_\alpha^{-1}(\overline{\Phi}(k)) - c\right],$$

which is equivalent to $\overline{\Phi}(k) = \Gamma_\alpha(c)$. This is equivalent to the allocation rule in Lemma 1 (see its proof for details). The second-order condition is satisfied whenever the first-order condition is satisfied if $\overline{\Phi}(v)$ is monotone. □

The remainder of the proof now proceeds in four steps: we show that (a) for the limiting distributions $\tilde{F}^*$ and $\tilde{G}^*$ the expectational fee ϖ^* is equal to the limiting fee $\frac{\alpha}{\alpha+\sigma}\tilde{p}$, (b) the expectational fee converges to the limiting fee, (c) the transaction fee $\tilde{\omega}$ is equal to the limiting fee for the limiting distributions, and (d) the transaction fee converges to $\frac{\alpha}{\alpha+\sigma}\tilde{p}$.

Step (a): First, we show that linearity of fees holds for the denormalized limiting distributions F^* and G^*. For simplicity, denote the supports of the denormalized limiting distributions as $[\underline{v}, \overline{v}]$ and $[\underline{c}, \overline{c}]$. The distributions are hence $F^*(v) = 1 - [(\overline{v}-v)/(\overline{v}-\underline{v})]^\beta$ and $G^*(c) = [(c-\underline{c})/(\overline{c}-\underline{c})]^\sigma$. The virtual cost function is linear: $\Gamma^*_\alpha(c) = c + (c-\underline{c})\alpha/\sigma$. The optimal expectational fees given in Lemma 2 can be rearranged to yield

$$\overline{\omega}^*(p) = p - E_v[\Gamma^{*-1}_\alpha(\overline{\Phi}^*(v))|v \geq p] = p - \Gamma^{*-1}_\alpha(E_v[\overline{\Phi}^*(v)|v \geq p]) = p - \Gamma^{*-1}_\alpha(p),$$

where the second equality stems from the linearity of Γ^*_α and the third from the well-known fact that for any p and any distribution $\overline{F}$ with virtual value $\overline{\Phi}$, $E_v[\overline{\Phi}(v)|v \geq p] = p$. Plugging in the functional form for Γ^*_α yields

$$\overline{\omega}^*(p) = (p - \underline{c}) \left[\frac{\alpha}{\alpha + \sigma}\right].$$

This implies that the equation for $\overline{\omega}^*$ in (6) holds for the limiting distributions F^* and G^*, because of the definitions of $\tilde{\omega}$ and $\tilde{p}$.

Step (b): Next, we show convergence to linearity. For this, it is useful to consider a linear transformation of the original problem, such that the length of the support is 1 for both F and G, and the lower bound is 0. This can be done without loss of generality. Formally, the support of the seller's distribution $[\underline{c}_j, (\overline{v}_j - \underline{c}_j)/u^S_j + \underline{c}_j]$ is transformed to $[0, 1]$ and the support of the buyer's distribution becomes $[\overline{v}_j - (\overline{v}_j - \underline{c}_j)/u^B_j, \overline{v}_j]$ to $[u_j - 1, u_j]$ with some $u_j > 0$. Note that as $j \to \infty$, $u_j \to 0$. In part of the following analysis, we will drop the subscript j and simply write $u \to 0$.

This has the advantage that the transformed distributions are only shifted and not stretched compared to F and G. Call these transformed distributions $\hat{F}_j$ and $\hat{G}_j$, with $\hat{G}_j(\hat{c}) = G(\hat{c})$ and $\hat{F}_j(\hat{v}) = F(\hat{v} + (1-u))$. The transformed fee is

$$\hat{\overline{\omega}}(\hat{p}) = u\hat{p} - \frac{\int_{\hat{p}}^1 \hat{\Gamma}^{-1}_\alpha(\hat{\overline{\Phi}}(u\hat{v}))d\hat{\overline{F}}(u\hat{v})}{1 - \hat{\overline{F}}(u\hat{p})} \tag{14}$$

where the expression comes from plugging in $u\hat{p}$ for p in the expression in Lemma 2.

We need to show that the expression in the integral uniformly converges to its limit, which implies convergence of the integral and also convergence of the whole expression for $\hat{\overline{\omega}}$.

By the definition of β we have

$$\lim_{u \to 0} \frac{\partial}{\partial(u\hat{v})} \left[\frac{1 - \hat{F}(u\hat{v})}{\hat{f}(u\hat{v})}\right] = \lim_{v' \to 1} \left[\frac{1 - F(v')}{f(v')}\right]' = \frac{1}{\beta}.$$

This implies that

$$\frac{1}{u}\left[\frac{1-\hat{F}(u\hat{v})}{f(u\hat{v})}\right] \overset{u\to 0}{\rightrightarrows} \frac{\hat{v}}{\beta}$$

and hence

$$\frac{1}{u}\hat{\Phi}(u\hat{v}) \overset{u\to 0}{\rightrightarrows} \hat{v} - \frac{1-\hat{v}}{\beta}$$

where the double arrow $\rightrightarrows$ stands for uniform convergence. By a similar logic

$$\frac{1}{u}\hat{\Gamma}_\alpha(u\hat{c}) \overset{u\to 0}{\rightrightarrows} \hat{c}\left(1+\frac{\alpha}{\sigma}\right)$$

and hence

$$\frac{1}{u}\hat{\Gamma}_\alpha^{-1}(ux) \overset{u\to 0}{\rightrightarrows} \frac{x}{1+\alpha/\sigma},$$

because uniform convergence of a function implies uniform convergence of its inverse (see for example Barvinek et al. (1991)).

Observe that

$$\hat{\overline{F}}(k) = \hat{\overline{F}}_\infty(\hat{R}^{-1}(k)), \qquad \hat{F}_\infty(p) = \frac{1-\hat{F}_{(1)}(p)}{1-\delta\hat{F}_{(1)}(p)}, \qquad \hat{F}_{(1)}(p) = \sum_{B=0}^{\infty} \pi_B \hat{F}(p)^B \tag{15}$$

and let

$$\hat{R}_j(p) = \frac{\int_p^{u_j} \hat{\Phi}(v)d\hat{F}_{(1)}(v)}{1-\hat{F}_{(1)}(p)}. \tag{16}$$

By Theorem 1 the expressions in (15) uniformly converge to their respective limits if $\hat{R}_j^{-1}$ uniformly converges. $\hat{R}_j^{-1}$ converges uniformly if $\hat{R}_j$ converges uniformly. So we are left to show that $\hat{R}_j$ converges uniformly in order to show uniform convergence of the integrand in (14) and hence convergence of $\hat{\overline{\omega}}$.

Since the integrand in the integral in $\hat{R}$ converges uniformly, $\hat{R}$ converges pointwise to its limit. Further, observe that the sequence $\hat{R}_j(p)$ with

$$\hat{R}_j(p) = \frac{\int_p^{u_j} \hat{\Phi}(v)d\hat{F}_{(1)}(v)}{1-\hat{F}_{(1)}(p)} = \frac{\int_{p+1-u_j}^{1} (\Phi(y)-(1-u_j))f_{(1)}(y)dy}{1-F_{(1)}(p+1-u_j)}$$

monotonically increases as j goes to infinity (and thus u_j goes to zero). Pointwise convergence and monotonicity of the sequence imply uniform convergence of $\hat{R}_j$ by Dini's theorem. Putting this together implies that $\hat{\overline{\omega}}$ converges to $\overline{\omega}^*$.

Step (c): Observe that if expectational fees are linear, the transaction fees are equal to expectational fees, since a linear function can be taken into an expectation. Hence the transaction fee $\omega(p) = \overline{\omega}(p)$ is also linear in the limit.

Step (d): Next, we turn to convergence of the normalized transaction fee $\hat{\omega}$.

From Proposition 2 and the arguments that precede it, we know that the optimal transaction fee $\omega(p)$ is given by

$$\omega(p) = p - \frac{\int_p^{\overline{v}} [\Gamma_\alpha^{-1}(\tilde{\Phi}(v)) + \delta V(\Gamma_\alpha^{-1}(\tilde{\Phi}(v)))] dF(v)}{1 - F(p)},$$

where $V(c) = \int_c^{\Gamma_\alpha^{-1}(\overline{v})} (1 - F_\infty(\tilde{\Phi}^{-1}(\Gamma_\alpha(y))))dy$. Defining $B := \int_p^{\overline{v}} \Gamma_\alpha^{-1}(\tilde{\Phi}(v))dF(v)$ and $A := \int_p^{\overline{v}} \int_{\Gamma_\alpha^{-1}(\tilde{\Phi}(p))}^{\Gamma_\alpha^{-1}(\overline{v})} (1 - F_\infty(\tilde{\Phi}^{-1}(\Gamma_\alpha(y))))dydF(v)$, we can thus write $\omega(p)$ as $\omega(p) = p - (B + \delta A)/(1 - F(p))$. Using $\tilde{\Phi}(p) = \overline{\Phi}(R(p))$ it is clear that the integrand in B converges uniformly and hence B converges. Reversing the order of integration in A and integrating we obtain

$$\begin{aligned} A &= \int_{\Gamma_\alpha^{-1}(p)}^{\Gamma_\alpha^{-1}(\overline{v})} \int_p^{\tilde{\Phi}(\Gamma_\alpha(y))} dF(v)(1 - F_\infty(\tilde{\Phi}^{-1}(\Gamma_\alpha(y))))dy \\ &= \int_{\Gamma_\alpha^{-1}(p)}^{\Gamma_\alpha^{-1}(\overline{v})} (F(\tilde{\Phi}(\Gamma_\alpha(y))) - F(p))(1 - F_\infty(\tilde{\Phi}^{-1}(\Gamma_\alpha(y))))dy. \end{aligned}$$

The integrand in the last expression is a combination of functions which we have shown to converge uniformly, hence we get convergence of A. Putting this together, we get that $\hat{\omega}$ converges to the limit given by substituting in the limiting distributions for F and G. □

Proof of Proposition 4. The Taylor expansion of $\Gamma_\alpha^{-1}(x)$ around $\overline{v}$ is

$$\Gamma_\alpha^{-1}(x) = \Gamma_\alpha^{-1}(\overline{v}) + [\Gamma_\alpha^{-1}(\overline{v})]'(x - \overline{v}) + \frac{[\Gamma_\alpha^{-1}(\overline{v})]''}{2}(x - \overline{v})^2 + \sum_{n=3}^{\infty} \frac{[\Gamma_\alpha^{-1}(\overline{v})]^{(n)}}{n!}(x - \overline{v})^n.$$

Denote the nth derivative at $\overline{v}$ as $\overline{\gamma}_n := [\Gamma_\alpha^{-1}(\overline{v})]^{(n)}$. We further use the shorthand

$\varphi := \Phi(v)$. The net price received by the seller can be rearranged as

$$\begin{aligned}
p - \omega(p) =& E[\Gamma_\alpha^{-1}(\varphi)|v \geq p] \\
=& \sum_{n=0}^{\infty} \frac{\overline{\gamma}_n}{n!} E[(\varphi - \overline{v})^n | v \geq p] \\
=& \sum_{n=0}^{\infty} \frac{\overline{\gamma}_n}{n!} \left\{ (E[\varphi|v \geq p] - \overline{v})^n - (E[\varphi|v \geq p] - \overline{v})^n + E[(\varphi - \overline{v})^n | v \geq p] \right\} \\
=& \Gamma_\alpha^{-1}(E[\varphi|v \geq p]) + \sum_{n=0}^{\infty} \frac{\overline{\gamma}_n}{n!} \left\{ E[(\varphi - \overline{v})^n | v \geq p] - (E[\varphi|v \geq p] - \overline{v})^n \right\} \\
=& \Gamma_\alpha^{-1}(p) + \sum_{n=0}^{\infty} \frac{\overline{\gamma}_n}{n!} \left\{ E[(\varphi - \overline{v})^n | v \geq p] - (E[\varphi|v \geq p] - \overline{v})^n \right\},
\end{aligned}$$

where the second equality stems from the Taylor expansion, the fourth from reversing a Taylor expansion, and the fifth from the fact that $E[\Phi(v)|v \geq p] = p$. Note that for $n = 0$ and $n = 1$, the expressions in curly braces cancel out in the last expression. For $n = 2$, the expression in curly braces is the conditional variance $\text{Var}[\varphi - \overline{v}|v \geq p] = E[(\varphi - \overline{v})^2|v \geq p] - (E[\varphi|v \geq p] - \overline{v})^2$. This completes the proof. □

A.2 Lemma 1

The proof Lemma 1 relies on mechanism design. We say that a *mechanism* is active in period t if the seller has not exited prior to t, which can happen because a transaction has occurred or because of the exogenously given probability $1 - \delta$ of dropping out from one period to the next. As mentioned, one can alternatively and equivalently interpret δ as the pure per period survival probability of the seller, or as a discount factor that reflects pure and common time preferences, or a as a combination of the survival probability and time preferences. However, the interpretation of many concepts used in the mechanism design framework is most straight forward if one interprets δ as a pure survival probability. After the seller exits, no good is left to be traded and the mechanism shuts down. The following, therefore, applies only to active mechanisms.

A mechanism is said to be a *direct mechanism* if it asks all agents who participate in the mechanism to report their types. For the seller, who is present at date 0, this simply means that he reports his cost c. A direct mechanism then asks all buyers who enter in period t to report their valuations $v_b \in [\underline{v}, \overline{v}]$ to the mechanism. The realization of the valuations of buyers who do not enter are set to $v_b = -\infty$. Let $\boldsymbol{v}_t = (v_b^t)_{b=1}^{\overline{B}}$ be a vector of such reports by buyers in period t with buyers label $b = 1, .., \overline{B}$ and let $\boldsymbol{v} = (\boldsymbol{v}_t)_{t=0}^{\infty}$ be a sequence of such reports.

A direct mechanism specifies the probability $Q_S^t(\boldsymbol{v}_t, c)$ that the seller sells in period t and the probability $Q_b^t(\boldsymbol{v}_t, c)$ that buyer b receives the good and the payment $M_S^t(\boldsymbol{v}_t, c)$ made from the mechanism to the seller and the payments made by buyers b to mechanism $M_b^t(\boldsymbol{v}_t, c)$, given reports $(\boldsymbol{v}_t, c)$ and given that the mechanism is still active.

Feasibility further requires

$$\sum_{b=1}^{\overline{B}} Q_b^t(\boldsymbol{v}_t, c) \leq Q_S^t(\boldsymbol{v}_t, c) \tag{17}$$

for all t and all $(\boldsymbol{v}_t, c)$. Accordingly, the mechanism ceases to be active in period t with probability $Q_S^t(\boldsymbol{v}_t, c)$, and it proceeds to period $t+1$ with probability $(1-\delta)(1-Q_S^t(\boldsymbol{v}_t, c))$.

Let $\boldsymbol{Q}_B^t(\boldsymbol{v}_t, c) = (Q_1^t(\boldsymbol{v}_t, c), .., Q_{\overline{B}}^t(\boldsymbol{v}_t, c))$ and $\boldsymbol{M}_B^t(\boldsymbol{v}_t, c) = (M_1^t(\boldsymbol{v}_t, c), .., M_{\overline{B}}^t(\boldsymbol{v}_t, c))$. For a given $(\boldsymbol{v}, c)$, let

$$\boldsymbol{Q}_S(\boldsymbol{v}, c) = \left(\boldsymbol{Q}_S^t(\boldsymbol{v}_t, c)\right)_{t=0}^{\infty} \quad \text{and} \quad \boldsymbol{Q}_B(\boldsymbol{v}, c) = \left(\boldsymbol{Q}_B^t(\boldsymbol{v}_t, c)\right)_{t=0}^{\infty}$$

and

$$\boldsymbol{M}_S(\boldsymbol{v}, c) = \left(M_S^t(\boldsymbol{v}_t, c)\right)_{t=0}^{\infty} \quad \text{and} \quad \boldsymbol{M}_B(\boldsymbol{v}, c) = \left(\boldsymbol{M}_B^t(\boldsymbol{v}_t, c)\right)_{t=0}^{\infty}.$$

Letting $\boldsymbol{Q}$ and $\boldsymbol{M}$ be, respectively, collections $\{\boldsymbol{Q}_S(\boldsymbol{v}, c), \boldsymbol{Q}_B(\boldsymbol{v}, c)\}$ and $\{\boldsymbol{M}_S(\boldsymbol{v}, c), \boldsymbol{M}_B(\boldsymbol{v}, c)\}$ for all possible $(\boldsymbol{v}, c)$, a direct mechanism is summarized by $\langle \boldsymbol{Q}, \boldsymbol{M} \rangle$ where $\boldsymbol{Q}$ satisfies (17). It is said to satisfy interim individual rationality and incentive compatibility if it satisfies these constraints for every possible type of every agent who participates at the period the agent first participates in the mechanism. For buyers, the latter condition is vacuously satisfied because they participate in the mechanism in one period only, if they participate at all. For the seller it means that these constraints have to be satisfied at date 0 only.

The focus on direct mechanisms is now easily seen to be without loss of generality: In every period t, no mechanism that respects buyers' incentive and interim individual rationality constraints can do better than a direct mechanism that respects these constraints (see e.g. Krishna, 2002). Applied iteratively, this then implies that no incentive compatible and interim individually rational mechanism can do better than an incentive compatible and interim individually rational mechanism that asks the seller to report his type in period 0.

The analysis is greatly simplified by using two concepts. The first is the *ultimate probability of selling* for a seller who reports type c

$$q_S(c) := E_{\boldsymbol{v}}\left[\sum_{t=0}^{\infty} Q_S^t(\boldsymbol{v}_t, c)\prod_{\tau=0}^{t-1}\delta(1 - Q_S^\tau(\boldsymbol{v}_\tau, c))\right],$$

which was introduced in Satterthwaite and Shneyerov (2008). We introduce a second, novel concept, the *ultimate conditional expected revenue*, which we will describe later.[49]

The seller's expected discounted payment is

$$m_S(c) := E_{\boldsymbol{v}}\left[\sum_{t=0}^{\infty} M_S^t(\boldsymbol{v}_t, c)\prod_{\tau=0}^{t-1}\delta(1 - Q_S^\tau(\boldsymbol{v}_\tau, c))\right].$$

In a direct mechanism, the seller of type c who reports truthfully has thus an expected discounted payoff of

$$\mathcal{W}_S(c) = m_S(c) - q_S(c)c,$$

while the intermediary's expected discounted payoff when facing a seller who reports to be of type c is

$$\mathcal{W}_I(c) = E_{\boldsymbol{v}}\left[\sum_{t=0}^{\infty}\left(\sum_{b=1}^{\overline{B}} M_b^t(\boldsymbol{v}_t, c)\right)\prod_{\tau=0}^{t-1}\delta(1 - Q_S^\tau(\boldsymbol{v}_\tau, c))\right] - m_S(c),$$

where the notation $\mathcal{W}_i(c)$ for $i = I, S$ emphasizes that we are referring to payoffs in a direct mechanism as opposed to fee-setting as defined in Section 2. The natural extension of the objective function (1) to the general mechanism design setup is then

$$\max_{\langle \boldsymbol{Q}, \boldsymbol{M}\rangle} E_c[\alpha\mathcal{W}_I(c) + (1-\alpha)(\mathcal{W}_I(c) + \mathcal{W}_S(c))] \tag{18}$$

subject to incentive compatibility and interim individual rationality constraints of buyers and the seller. As there is no other restriction on the mechanisms used, this objective

[49] Satterthwaite and Shneyerov (2008) did not need this second concept, since their analysis is mostly about a full-trade equilibrium, in which all sellers trade with probability 1.

is more general than (1), which is confined to fee-setting. However, as we will show, the objective in (18) can be maximized with an appropriately chosen sequence of transaction fees $\boldsymbol{\omega}$. Moreover, we will show that individual rationality constraints are not only satisfied in the interim stage but also *ex post* and that the seller's incentive constraint can be satisfied period by period. While the results concerning *ex post* individual rationality of buyers is immediate because of the nature of second-price auctions, it is far from obvious *a priori* that such a mechanism exists in the dynamic setup with two-sided private information and arbitrary α we study.[50]

Standard arguments imply that a direct mechanism is incentive compatible for the seller if and only if it such that $q_S(c)$ is monotone in c and that in any direct, incentive compatible mechanism

$$m_S(c) = q_S(c)c + \int_c^{\overline{c}} q_S(x)dc + \mathcal{W}_S(\overline{c}) \tag{19}$$

holds (see e.g. Krishna, 2002). Monotonicity of $q_S(c)$ implies that the interim individual rationality constraint will be satisfied if it is satisfied for the seller of type $\overline{c}$, that is if $\mathcal{W}_S(\overline{c}) \geq 0$ (and if the seller's incentive constraint is satisfied). Because $\mathcal{W}_S(\overline{c})$ enters the objective function as the constant $-\alpha\mathcal{W}_S(\overline{c})$, it will be optimal to set $\mathcal{W}_S(\overline{c}) = 0$ for any $\alpha \in [0, 1]$.

We say that the good is auctioned off in period t with reserve p_t if $Q_b^t(\boldsymbol{v}_t, c) = 1$ if $v_b = \max\{\boldsymbol{v}_t\}$ and $v_b \geq p_t$ and $Q_i^t(\boldsymbol{v}_t, c) = 0$ for all $i = 1, .., \overline{B}$ otherwise.[51]

Lemma 3. *A mechanism $\langle \boldsymbol{Q}, \boldsymbol{M} \rangle$ is optimal only if the good is auctioned off in every period t at some reserve p_t.*

Proof of Lemma 3. Suppose to the contrary that the optimal mechanism, denoted $\langle \hat{\boldsymbol{Q}}, \hat{\boldsymbol{M}} \rangle$, does not auction off the good at some reserve in period t and in some states $(\boldsymbol{v}_t, c)$. This implies that with positive probability the good is sold in period t to a buyer whose valuation is not the highest amongst all the buyers present. Consider then an alternative

[50] The direct mechanism problem that we set up here is thus a relaxed problem, and we will show that the additional constraints are not binding. For an analysis of *ex post* individual rationality of a bilateral trade problem where the intermediary makes zero profit, see Gresik (1991).

[51] This definition neglects the possibility of ties at the highest value, which have probability 0. If one wants to account for such ties explicitly, one can arbitrarily set $Q_b^t(\boldsymbol{v}_t, c) = 1$ for the buyer b with the highest valuation and, say, the highest index b amongst all buyers with the highest value.

mechanism that coincides with $\langle \hat{\boldsymbol{Q}}, \hat{\boldsymbol{M}} \rangle$ except for the states in period t in which $\langle \hat{\boldsymbol{Q}}, \hat{\boldsymbol{M}} \rangle$ does not auction off the good. Let the alternative mechanism sell the good to the highest value buyer in all those instances for which $\langle \hat{\boldsymbol{Q}}, \hat{\boldsymbol{M}} \rangle$ sells it to some other buyer. This alternative mechanism will increase the broker's payoff $\mathcal{W}_I(c)$ by increasing the revenue it raises while leaving the seller's payoff $\mathcal{W}_S(c)$ unaffected. Therefore, the mechanism $\langle \hat{\boldsymbol{Q}}, \hat{\boldsymbol{M}} \rangle$ cannot be optimal. □

Lemma 3 implies that the choice set for allocation rules $\boldsymbol{Q}$ can be narrowed down to sequences of reserves $\boldsymbol{p}(c) = (p_t(c))_{t=0}^{\infty}$, one sequence for every seller type c, with the understanding that, provided the mechanism is still active in period t, the good will be sold to the buyer with the highest valuation present in that period, provided this valuation is no less than $p_t(c)$. Letting $k_t := R(p_t)$ denote the expected transaction price in period t, conditional on a transaction occurring in period t, choosing a sequence of reserves $\boldsymbol{p}(c)$ is equivalent to choosing a sequence $\boldsymbol{k}(c) = (k_t(c))_{t=0}^{\infty}$ of expected transaction prices (conditional on a transaction occurring).

The key observations are the following. For any sequence of expected transaction prices $\boldsymbol{k} = (k_t)_{t=0}^{\infty}$, let

$$q_t(\boldsymbol{k}) := \big(1 - F_{(1)}(R^{-1}(k_t))\big) \prod_{\tau=0}^{t-1} \delta F_{(1)}(R^{-1}(k_\tau))$$

denote the discount factor, adjusted for the probability of prior sale (and for the probability of prior exit if δ is interpreted as probability of survival), for a transaction occurring in period t. The number $\sum_{t=0}^{\infty} q_t(\boldsymbol{k}) k_t$ is then the expected discounted transaction price, or average price for short, given $\boldsymbol{k}$ while $\sum_{t=0}^{\infty} q_t(\boldsymbol{k})$ is the ultimate probability of selling. The number

$$k = \frac{\sum_{t=0}^{\infty} q_t(\boldsymbol{k}) k_t}{\sum_{t=0}^{\infty} q_t(\boldsymbol{k})} \tag{20}$$

has then the interpretation of an *ultimate conditional expected revenue.* The simplest interpretation for k can be provided if one interprets $1 - \delta$ purely as the probability of dropping out from one period to the next (without any impatience on top of that). With this interpretation, k is the expected revenue conditional on trading and not dropping out. If impatience stems from a time preference rather than a drop-out probability,

k cannot be simply interpreted as a conditional expectation, but should be viewed as an abstract mathematical concept that helps to unite probabilities and discounting and simplifies calculations.[52]

A mechanism can only be optimal if it maximizes the ultimate conditional expected revenue $(\sum_{t=0}^{\infty} q_t(\boldsymbol{k})k_t)/(\sum_{t=0}^{\infty} q_t(\boldsymbol{k}))$ for a given ultimate probability of selling $\sum_{t=0}^{\infty} q_t(\boldsymbol{k})$. By a duality argument, it also holds that a mechanism can only be optimal if it maximizes the ultimate probability of selling for a given ultimate conditional expected revenue.

Proof of Lemma 1. For $k \in [\underline{v}, \overline{v}]$ let

$$1 - \overline{F}_T(k) := \max_{(k_t)_{t=0}^T} \left\{ \sum_{t=0}^{T} q_t(\boldsymbol{k}) \right\} \quad \text{s.t.} \quad \frac{\sum_{t=0}^{T} q_t(\boldsymbol{k})k_t}{\sum_{t=0}^{T} q_t(\boldsymbol{k})} = k,$$

and define

$$1 - \overline{F}(k) := \lim_{T\to\infty} 1 - \overline{F}_T(k).$$

Let $(k_t^*(k))_{t=0}^T$ be a maximizer of $\max_{(k_t)_{t=0}^T} \left\{ \sum_{t=0}^{T} q_t(\boldsymbol{k}) \right\}$ s.t. $\frac{\sum_{t=0}^{T} q_t(\boldsymbol{k})k_t}{\sum_{t=0}^{T} q_t(\boldsymbol{k})} = k$. Under stationarity, we have $k_t^*(k) = k$ for all t. Therefore,

$$\begin{aligned} 1 - \overline{F}(k) &= \lim_{T\to\infty} \sum_{t=0}^{T} \left(\prod_{\tau=0}^{t-1} \delta F_{(1)}(R^{-1}(k)) \right) \left(1 - F_{(1)}(R^{-1}(k))\right) \\ &= \lim_{T\to\infty} \frac{1 - \delta^{T+1} F_{(1)}(R^{-1}(k))^{T+1}}{1 - \delta F_{(1)}(R^{-1}(k))} \left(1 - F_{(1)}(R^{-1}(k))\right) \\ &= \frac{1 - F_{(1)}(R^{-1}(k))}{1 - \delta F_{(1)}(R^{-1}(k))} = 1 - F_\infty(R^{-1}(k)). \end{aligned}$$

Therefore, for a given c and k, we now have $\mathcal{W}_I(c) = k(c)(1 - \overline{F}(k(c))) - m_S(c)$ and $\mathcal{W}_S(c) = m_S(c) - q_S(c)c$. Using incentive compatibility (19) and $\mathcal{W}_S(\overline{c}) = 0$ by individual rationality, the objective given c becomes

$$\alpha \mathcal{W}_I(c) + (1-\alpha)(\mathcal{W}_I(c) + \mathcal{W}_S(c)) = k(1 - \overline{F}(k)) - c q_S(c) - \alpha \int_c^{\overline{c}} q_S(x) dx.$$

Substituting $q_S(k) = 1 - \overline{F}(k)$ and integrating after reversing the order of integration in the double-integral then yields the objective function

$$\max_{k(c)} \int_{\underline{c}}^{\overline{c}} \left[k(c) - \Gamma_\alpha(c)\right] (1 - \overline{F}(k(c))) g(c) dc \tag{21}$$

[52]One of the advantages of using the ultimate conditional expected revenue is that it avoids the problem of a standard conditional expected revenue with a time preference interpretation of discounting: for any positive constant per period probability of sale, the seller eventually trades with probability 1, so that conditioning on trade occurring would not be a useful concept.

with

$$\Gamma_\alpha(c) := c + \alpha \frac{G(c)}{g(c)}.$$

Observe that monotonicity of $\Gamma(c)$ implies monotonicity of $\Gamma_\alpha(c)$. The integral can be maximized pointwise by choosing k such that

$$0 = -\overline{f}(k(c)) \left[\overline{\Phi}(k(c)) - \Gamma_\alpha(c)\right],$$

which is equivalent to $k(c) = \overline{\Phi}^{-1}(\Gamma_\alpha(c))$. This is a monotone function and thus incentive compatible. Moreover, the second-order condition for a maximum is satisfied whenever the first-order condition is satisfied if $\overline{\Phi}(v)$ is monotone.

This means that the optimal allocation rule is such that trade takes place as soon as

$$\overline{\Phi}(k) \geq \Gamma_\alpha(c). \tag{22}$$

Let $k^*(c) := \overline{\Phi}^{-1}(\Gamma_\alpha(c))$. Since $\overline{F}(k) = F_\infty(R^{-1}(k))$, $\overline{\Phi}(k) = \tilde{\Phi}(R^{-1}(k))$. According to (22) trade should take place as soon as $v \geq R^{-1}(k^*(c))$, which because of the afore-noted equalities and the monotonicity of $\tilde{\Phi}$, is equivalent to $\tilde{\Phi}(v) \geq \Gamma_\alpha(c)$ as claimed in the lemma. □

B Extension: Non-Stationarity

Assume now that the environment is described by known sequences $\boldsymbol{\delta} = (\delta_t)_{t=0}^\infty$ and $\boldsymbol{F} = (F_t)_{t=0}^\infty$, where δ_t is the discount factor in period t and F_t is the distribution from which buyers' types are drawn in period t and that for all t, $\Phi_t(v) = v - \frac{1-F_t(v)}{f_t(v)}$ is monotone in v. One example is the exponential discounting (or constant drop out probability) $\delta_\tau = \delta$ considered so far. Another is exponential discounting up to a deadline T after which the seller leaves the market for sure ($\delta_\tau = \delta$ for $\tau \leq T$ and $\delta_\tau = 0$ for $\tau > T$). Let π_B^t describe the arrival process of buyers in period t and denote by $F_{(1),t}(v)$ and $F_{(2),t}(v)$ the distributions of the highest and second-highest draw in t. Expected revenue given reserve p in period t, conditional on trade in period t, is then given as $R_t(p) = \frac{\int_p^{\overline{v}} \Phi_t(v) dF_{(1),t}(v)}{1-F_{(1),t}(p)}$.

Given a sequence $\boldsymbol{k}$ of expected transaction prices conditional on trade, the seller's ultimate probability of selling is still given as $\sum_{t=0}^{\infty} q_t(\boldsymbol{k})$, where

$$q_t(\boldsymbol{k}) := \left(1 - F_{(1),t}(R_t^{-1}(k_t))\right) \prod_{\tau=0}^{t-1} \delta_\tau F_{(1),\tau}(R_\tau^{-1}(k_\tau)).$$

Next define $1-\overline{F}(k) := \lim_{T\to\infty} 1-\overline{F}_T(k)$, where $1-\overline{F}_T(k)$ is the maximum of $\sum_{t=1}^{T} q_t(\boldsymbol{k})$ subject to the constraint $(\sum_{t=0}^{T} q_t(\boldsymbol{k})k_t)/(\sum_{t=0}^{T} q_t(\boldsymbol{k})) = k$, as defined in the proof of Proposition 2. At date 0, the objective function that accounts for incentive compatibility provided the pointwise maximizer $k(c)$ of the integrand is monotone is then still given by (21), yielding the allocation rule allocation rule (22). Consequently, the functional form of the expectational fees $\overline{\omega}(k)$ under non-stationarity will be the same as under stationarity. Hence, it is as given in Lemma 2 in the proof of Proposition 3. This also implies that in the limit, as G converges to a mirrored Generalized Pareto distribution, the optimal expectational fee will be linear as in the stationary case.

Although $\omega_t(p)$ will in general vary over time because the environment is non-stationary, the linearity of the expectational fees $\overline{\omega}$ in the limit implies that the optimal transaction fees will be linear in the limit too.

C Extreme Value Theory

Extreme Value Theory For the convenience of the reader, this appendix provides a summary of the results of the theory of exceedences in extreme value theory that are the most important ones for the purposes of our paper. This summary is the content of Theorem 1 below. The theorem says that for any F that satisfies some weak regularity condition,

$$\lim_{u\to 0} 1 - \frac{1 - F(\overline{v} - u(\overline{v} - v))}{1 - F(\overline{v} - u(\overline{v} - \underline{v}))} = 1 - \left(\frac{\overline{v} - v}{\overline{v} - \underline{v}}\right)^{\beta} =: F^*(v), \tag{23}$$

where convergence is uniform and β is some constant. The left-hand side of (23) is the rescaled distribution conditional on being above the threshold $\overline{v} - u(\overline{v} - \underline{v})$. According to Theorem 1, this truncated and rescaled distribution converges to a Generalized Pareto distribution F^* as the threshold $\overline{v} - u(\overline{v} - \underline{v})$ goes to the finite upper bound $\overline{v}$.

The motivation for this theory was the empirical regularity found in many situations that the upper tail of a distribution is well approximated by a (Generalized) Pareto distribution. A prominent example is the distribution of the highest 20 percent of income and wealth in many countries, which was first observed by Vilfredo Pareto.[53] The theory of exceedences within extreme value theory deals with the distribution of a random variable conditional on being above a high threshold (for the original articles see Balkema and De Haan (1974), Pickands (1975); for a textbook see Falk et al. (2010)).

The general principle is described by the Pickands-Balkema-de Haan theorem (also called the second theorem of extreme value theory). For expositional simplicity, we provide a simplified version of the theorem, which is sufficient for our purposes. See Pickands (1975, Theorem 7) and Balkema and De Haan (1974) for the theorem itself. The theorem establishes a connection between the behavior of the maximum of a distribution and its upper tail. The relevant concept for the maximum is the domain of attraction:

Definition 1. *A distribution F is in the* domain of attraction of an extreme value distribution *if there exists a sequence of constants $a_n > 0$ and b_n real for $n = 1, 2, ...,$ such that*

$$\lim_{n\to\infty} [F(a_n x + b_n)]^n = F_{max}(x)$$

for every continuity point x of F_{max} for some non-degenerate distribution function F_{max} (see De Haan and Ferreira, 2006, p. 4).

This means that for n independently and identically distributed random variables, $(\max\{X_1, X_2, ..., X_n\} - b_n)/a_n$ has a non-degenerate distribution as n goes to infinity.

The following theorem holds.

Theorem 1. *(Simplified version of the Pickands-Balkema-de Haan Theorem) Assume F has a finite upper bound and $f(v) > 0$ for all $v \in (\underline{v}, \overline{v})$. Then F has a Generalized Pareto upper tail, formally*

$$\lim_{u\to 0} 1 - \frac{1 - F(\overline{v} - u(\overline{v} - v))}{1 - F(\overline{v} - u(\overline{v} - \underline{v}))} = 1 - \left(\frac{\overline{v} - v}{\overline{v} - \underline{v}}\right)^{\beta}, \tag{24}$$

[53]Other examples include the distribution of the strength of earthquakes in historical data (which tend to contain only the most severe earthquakes); and for the discrete type variant of the Pareto distribution – Zipf's law – the distribution of the frequency of the most common words in a larger text and the sizes of the largest cities in most countries.

for some constant β, where convergence is uniform, if and only if F is in the domain of attraction of an extreme value distribution.

The left-hand side of (24) is the rescaled distribution conditional on being above the threshold $\overline{v} - u(\overline{v} - \underline{v})$. The right-hand side is the cumulative distribution function of a finite upper bound Generalized Pareto distribution.

Proof of Theorem 1. See Theorem 7 in Pickands (1975). Note that for our setup ($\overline{v}$ finite and $f(v) > 0$ for all $v \in (\underline{v}, \overline{v})$) the definition of F having a Generalized Pareto upper tail given in Definition 4 in Pickands (1975) simplifies to (24). □

The literature on extreme value theory states several sufficient conditions for a distribution to be in the domain of attraction of an extreme value distribution. We state the one most suitable for our purposes.

Theorem 2. *Assume F has a finite upper bound. F is in the domain of attraction of an extreme value distribution if the von Mises condition*

$$\lim_{v \to \overline{v}} \frac{d}{dv} \left[\frac{1 - F(v)}{f(v)} \right] = \overline{\beta}, \tag{25}$$

for some constant $\overline{\beta}$, holds.

Proof. See, for example, Theorem 1.1.8 in De Haan and Ferreira (2006, p. 15). □

As stated in the literature, even this sufficient condition is weak and is satisfied by all "textbook" continuous distributions, such as uniform, Beta, bounded Generalized Pareto, inverse Weibull and (for the infinite upper bound counterpart of the condition) the normal, exponential, Cauchy, and infinite upper bound Generalized Pareto distribution.

Often, the Generalized Pareto distribution is defined with the parametrization

$$F^*(v) = 1 - \left(1 + \frac{\xi(v - \mu)}{\sigma} \right)^{-1/\xi}.$$

For $\xi < 0$ the distribution has a finite upper bound and corresponds to the parametrization used in this paper with $\underline{v} = \mu$, $\overline{v} = \mu - \sigma/\xi$, and $\beta = -1/\xi$. For $\xi \geq 0$, it has an infinite upper bound and lower bound μ. One obtains the exponential distribution as a special case as $\lim_{\xi \to 0} F^*(v) = 1 - e^{-(v-\mu)/\sigma}$. For $\xi > 0$ and $\sigma = \mu\xi$ one obtains the classical Type I Pareto distribution $F(v) = 1 - (\mu/v)^{1/\xi}$. For $\xi > 0$ one obtains the Type II Pareto distribution.

For infinite upper bounds, convergence can be stated as

$$\left(1 - \frac{1 - F(u + x)}{1 - F(u)}\right) - F_u^*(x) \overset{u \to \infty}{\to} 0,$$

for some Generalized Pareto distribution F_u^*. See the above mentioned references for more details.

Note that the characteristic property of Generalized Pareto distributions is that the inverse hazard rate is linear: $[(1 - F(v))/f(v)]' = \xi$. The special cases can be seen as the inverse hazard rate decreasing (bounded Generalized Pareto distribution), constant (exponential distribution), and increasing ((Non-Generalized) Pareto distribution). $\xi < 0$ corresponds to the common monotone hazard rate condition (that is, $f(v)/(1 - F(v))$ is increasing). $\xi < 1$ corresponds to Myerson's regularity condition $\Phi'(v) > 0$ and is also necessary to ensure that the distribution has a finite mean.