# Consumer Inattention and Bill-Shock Regulation

MICHAEL D. GRUBB

*Boston College*

*First version received July 2012; final version accepted May 2014 (Eds.)*

**Abstract**

For many goods and services such as electricity, healthcare, cellular-phone service, debit-card transactions, or those sold with loyalty discounts, the price of the next unit of service depends on past usage. As a result, consumers who are inattentive to their past usage but are aware of contract terms may remain uncertain about the price of the next unit. I develop a model of inattentive consumption, derive equilibrium pricing when consumers are inattentive, and evaluate bill-shock regulation requiring firms to disclose information that substitutes for attention. When inattentive consumers are sophisticated but heterogeneous in their expected demand, bill-shock regulation reduces social welfare in fairly-competitive markets, which may be the effect of the FCC's recent bill-shock agreement. If some consumers are attentive while others naively fail to anticipate their own inattention, however, then bill-shock regulation increases social welfare and can benefit consumers. Hence requiring zero-balance alerts in addition to the Federal Reserve's new opt-in rule for debit-card overdraft protection may benefit consumers.

*Key words*: inattention, bill-shock, consumer protection, penalty fee, loyalty discount, overconfidence, nonlinear pricing, imperfect recall, cellular, overdraft, naivete

*JEL Codes*: D18, D43, D82, L11, L13

## 1 Introduction

For many goods and services, including electricity, healthcare, cellular-phone service, debit-card transactions, and those sold with loyalty discounts, the price of the next unit of service depends on past usage. In many cases, such as electricity, cellular-phone service, and debit-card transactions, marginal prices increase sharply when consumers exceed specified usage thresholds. Consumers commonly cross such usage thresholds and accrue high fees without realizing it, resulting in *bill shock*, because they are inattentive and do not keep track of past usage.[1] For example, a cellular-phone user may not realize that the current call is charged a penalty (or *overage*) rate of 45 cents

[1]Bill shock also arises without inattention when multiple family members consume from the same family-talk plan or joint checking-account but do not continually update each other about purchases. Shared usage will be an alternative interpretation to inattention throughout the paper.

per minute, because he does not know that he has already used up his 500 included minutes. Similarly, a checking account holder may be unaware that her next debit transaction will incur a $35 overdraft penalty because she does not realize her checking balance is negative.

In other cases, such as airline loyalty programs, purchasing the flight which crosses a 25,000 mile threshold is effectively discounted by the value of the elite status perks it earns for the following year. In this case there is no bill shock but there may instead be a welcome surprise when an elite status card arrives in the mail if the customer has not been tracking her mileage. In still other cases, such as with the Medicare Part D "doughnut hole", marginal prices may first rise and then fall again at different usage thresholds all of which can be a surprise to cross without careful attention to cumulative annual prescription drug spending.

This paper's first goal, an important first step to understanding any of these markets, is to develop a model of consumption by inattentive consumers who do not keep track of their past usage. I will assume that: (1) Once a consumer signs a contract, two consumption opportunities arise sequentially. (2) Inattentive consumers make each decision to consume an additional unit without any recollection of prior usage. (Where important, I explicitly model inattention as a rational choice to avoid attention costs.) Given these two assumptions, I show that an inattentive consumer's optimal strategy is a threshold rule: to buy only those units valued above their expected marginal price.[2] Sophisticated consumers who anticipate their own inattention follow this strategy. Naive consumers who fail to anticipate their own inattention deviate from this strategy because they mispredict their future choices. Moreover, naively inattentive consumers overestimate the value of offered contracts because they underestimate the probability of paying surprise penalty fees (via endogenous overconfidence) and overestimate the probability of redeeming loyalty discounts or perks (via endogenous underconfidence).

This paper's second goal is to determine whether bill-shock regulation requiring firms to disclose information about transaction prices at the point of sale will benefit consumers or raise total welfare. (In the context of my model, such disclosures are a perfect substitute for attention.) Holding pricing fixed, bill-shock regulation should weakly benefit consumers, by giving them more information to make better choices. Presumably this is why bill-shock regulation has strong support from consumer groups and regulators.[3] Importantly, however, the focus of the analysis is on understanding the net

[2]This provides a micro-foundation for the threshold labor supply rule used by Saez (2002) and the consumption rules used by Borenstein (2009) and Grubb and Osborne (forthcoming). These papers use the threshold rules in demand or labor supply estimation, while I explore the supply-side ramifications of such behavior.

[3]FCC Chairman Julius Genachowski said, "something is clearly wrong with a system that makes it possible for consumers to run up big bills without knowing it," and a variety of consumer advocacy groups agree (Genachowski,

effect of such regulation when taking into account how firms will adjust prices in response to the regulation. As a result, the paper focuses on markets with unregulated prices such as the cellular phone and debit card examples rather than electricity or Medicare Part D.

In both cellular phone and debit card examples, firms have the ability to disclose whether a penalty fee is applicable at the point of sale. (Absent such disclosure I refer to the penalties as *surprise* penalty fees.) A mobile phone screen could flash "overage rate applies" before a call is made and a debit-card terminal could ask "Overdraft fee applies. Continue - Yes/No?" before processing transactions on an overdrawn account. That firms do not make these disclosures and oppose[4] regulation requiring such disclosure suggests that firms benefit from bill shock. In contrast, consumer groups and regulators such as the Federal Communications Commission (FCC) and the Federal Reserve Board believe that the lack of transparency is bad for consumers and bad for welfare, which has led to new regulation. For instance, in late 2011, US President Barack Obama said that,

> Far too many Americans know what it's like to open up their cell-phone bill and be shocked by hundreds or even thousands of dollars in unexpected fees and charges. But we can put an end to that with a simple step: an alert warning consumers that they're about to hit their limit before fees and charges add up.

Obama's statement was made at the announcement of an agreement between cellular carriers and the FCC to begin providing such usage alerts by April 2013 (CTIA - The Wireless Association, 2011). The Federal Reserve Board has been similarly concerned about overdraft fees on ATM and one-time debit-card transactions, fees which totalled $20 billion in 2009 (Martin, 2010). In response, since 2010, Federal Reserve Board rules prohibit such overdraft fees unless a consumer opts in to overdraft service. Moreover, the Consumer Financial Protection Bureau (CFPB) is currently considering additional overdraft fee regulation, and one option the CFPB could consider is to require banks to issue zero-balance alerts.

As stated above, this paper's second goal is to determine whether bill-shock regulation will benefit consumers or raise total welfare. To understand the effect of bill-shock regulation, however, I also answer a related question: Why do firms both charge *penalty fees* (so that high usage triggers

---

2010; Deloney, Sherry, Grant, Desai, Riley, Wood, Breyault, Gonzalez and Lennett, 2011).

[4]Prior to their recent bill-shock agreement with the FCC, the wireless industry trade group, C.T.I.A. - The Wireless Association, argued that the FCC's proposed bill-shock regulation "violates carriers' First Amendment protections. . . . against government compelled speech" (Altschul, Guttman-McCabe and Josef, 2011). Similarly, banks opposed the Federal Reserve Board's opt-in rule for overdraft protection (Federal Reserve Board, 2009).

high marginal charges) and make them a *surprise* by not alerting consumers when they cross the relevant threshold? I show that the answers to both questions depend on factors including (1) consumer sophistication, (2) consumer heterogeneity, and (3) market power.

Section 3 introduces the benchmark model with two benchmark assumptions to be relaxed later: (1) All consumers are inattentive and sophisticated (so anticipate their own inattention). (2) While consumers may have heterogeneous brand preferences, they all have the same demand (so there is no scope for volume-based price discrimination). The benchmark result is that firms have no incentive to charge surprise penalty fees and regulation will not affect firm profits or consumer surplus. This result is straightforward because, in the benchmark environment, a firm cannot do better than induce first-best consumption via marginal-cost pricing.

Moving beyond this initial benchmark, I develop three main results about equilibrium pricing and bill-shock regulation: (1) When consumers naively fail to anticipate their inattention, firms find it strictly optimal to charge surprise penalty fees or loyalty discounts. Bill-shock regulation restores marginal cost pricing. (2) If some consumers are attentive but more are naively inattentive, then bill-shock regulation increases social welfare, benefits naive consumers by ending their cross-subsidization of attentive consumers, and can benefit all consumers by increasing market competition. (3) In contrast, bill-shock regulation can lower social welfare and harm low-demand consumers when firms use multiple contracts and surprise penalty fees to price discriminate between low and high demand consumers who are sophisticated in their inattention. The result always holds in *fairly-competitive* markets (those that are highly but not perfectly competitive) because surprise penalty fees mitigate allocative distortions otherwise inherent in second-degree price discrimination. (Within fairly-competitive markets, the magnitude of the effect may be small and diminishes to zero as competition becomes perfect.)

These results suggest that the CFPB should consider requiring banks to issue low-balance alerts.[5] Banks price discriminate but typically do not vary overdraft fees across checking accounts to do so, as would be required if the fees were used to influence consumers' account choices. Thus the third main result does not apply to overdraft fees. However, banks' success in having 75% of consumers actively opt-in to overdraft protection programs is consistent with consumer naivete (Benoit, 2010). It is consistent because a consumer who believes she will be attentive prefers to opt-in.[6] Moreover, Stango and Zinman (2014) find that 48% of consumers pay no overdraft fees,

[5] Armstrong and Vickers (2012) discuss four overdraft fee regulations including low-balance alerts, opt-in or opt-out rules, fee caps, and salience requirements.

[6] An attentive consumer never overdrafts accidentally and hence never benefits if her bank declines an overdraft transaction. Thus there is no downside to opting in. There is a benefit to opting in, however, as long as there is a

consistent with heterogeneity in attention and substantial cross-subsidization across consumers. Thus the first and second main results suggest that low-balance alerts would be socially beneficial, protect naive consumers, and possibly benefit all checking account holders by stiffening competition.

Unfortunately, an assessment of the FCC's recent bill-shock agreement is less clear cut. In contrast to checking fees, cellular companies' overage charges are clearly used to help sort consumers into low and high included-minute contracts. However, one cannot immediately conclude from the third main result that the FCC's bill-shock agreement will reduce social welfare and harm consumers on inexpensive plans because the result assumes a fairly-competitive market and sophisticated consumers, whereas in fact cellular carriers do have substantial market power (Department of Justice, 2011) and evidence shows that their consumers' usage forecasts are overconfident (Grubb, 2009; Grubb and Osborne, forthcoming), consistent with naivete. Thus insights from all three main results are relevant but none give a clear prediction in this setting. Nevertheless, the results show that the FCC's recent bill-shock agreement could lower welfare and harm some consumers by a small amount, contrary to expectations of the FCC and consumer advocacy groups.[7]

Section 4 develops the first main result. It varies the benchmark model by assuming that consumers naively fail to anticipate their own inattention. Such consumers underestimate the chance of paying surprise penalty fees and overestimate the chance of collecting loyalty discounts. In this case, firms either charge surprise penalty fees or offer surprise loyalty discounts to exploit consumers' naivete. Consumers are assumed to be rationally inattentive, trading benefits of attention off against a cost $k$ of being attentive, thereby endogenously limiting the size of penalty fees or loyalty discounts. Naivete is rendered irrelevant by bill-shock regulation so that it cannot be exploited and marginal cost pricing is restored. Section 4.4 develops the second main result by enriching the model to include both attentive and naively inattentive consumers. In this case, bill-shock regulation ends cross-subsidization of attentive consumers by naively inattentive consumers. It also eliminates the socially inefficient distortions inherent in choices made by attentive consumers to avoid penalty fees and collect loyalty discounts. Moreover, if attentive consumers are more price sensitive than the naively inattentive (which is reasonable if attentive consumers are those who have both the time to track their usage and the time to comparison shop) then bill-shock regulation can benefit all consumers by stiffening competition. This can happen because bill-shock regulation

positive probability that the consumer values completing an overdraft transaction more than the overdraft fee.

[7]Complementary empirical work estimates calling demand using cellular billing data and simulates the effect of bill-shock regulation: Grubb and Osborne (forthcoming) predict that bill-shock regulation will lower social surplus by $26 per consumer and consumer surplus by $33 per consumer annually. In contrast, Jiang (2013) predicts that bill-shock regulation will raise social and consumer surplus by $24 per household annually.

eliminates an adverse selection problem (similar to that studied by Ellison (2005)) whereby the segment most attracted by a price cut (the attentive) is also the least profitable.

Section 5 develops the third main result. It enriches the benchmark model in a second direction by incorporating two ex ante types, with low and high expectations of future demand. Given such heterogeneity, the surprising result is that the combination of surprise penalty fees and consumer inattention can be socially valuable (as well as privately valuable to firms) and benefit low-demand consumers. Surprise penalty fees can be socially valuable because they can reduce allocative distortions imposed by price discriminating firms. Thus bill-shock regulation can lower social welfare and harm low-demand consumers. Moreover, this is always the case in fairly-competitive markets, although the magnitude of the effect may be small. The intuition is that, when consumers are inattentive, both surprise penalty fees and quantity distortions are useful tools for price discrimination because both relax incentive constraints. By substituting for attention and eliminating surprise, bill-shock regulation removes surprise penalty fees from the price-discrimination toolbox. This limits firms' ability to price discriminate, explaining their aversion to the regulation. Moreover, if surprise penalty fees and quantity distortions are substitutes, regulation leads to an increase in quantity distortions and reduces social welfare. In contrast, if the two tools are complements, regulation has the opposite effect. While, in general, surprise penalty fees and quantity distortions could be either substitutes or complements, they are always substitutes in fairly-competitive markets. This follows because (1) competition limits the additional markup firms would like to charge high-demand consumers; and (2) surprise penalty fees render quantity distortion unnecessary if differences in markups are small.

Proofs are in the appendix and the online appendix discusses the FCC's bill-shock agreement and the Federal Reserve Board's overdraft opt-in regulation in greater depth.

## 2 Evidence of inattention

US students appear inattentive to their past usage of cellular-phone minutes, delaying calls until off-peak periods even when many included peak minutes are unused at the end of a billing cycle (Grubb and Osborne, forthcoming). In contrast, Chinese phone users are attentive to past usage, which may be due to higher financial stakes (Yao, Mela, Chiang and Chen, 2012). Survey and transaction data show US and UK checking-account holders incur overdraft fees because they are inattentive to account balances (Stango and Zinman, 2009, 2014; Armstrong and Vickers, 2012). Stango and Zinman (2014) find that over 50 percent of US overdraft fees are avoidable by using alternative accounts with available liquidity and that "60% of overdrafters reported overdrafting

because they 'thought there was enough money in my account.'" Moreover, Stango and Zinman (2014) document prima facie evidence of heterogeneity in consumer attention levels to account balances. They find both heterogeneity in overdraft fee payments (48% of checking account holders never pay an overdraft fee while the remaining 52% do so on average every 3 months) and that those who do not pay overdraft fees nevertheless come within $100 of doing so in most months (suggesting fees were avoided through attention rather than a large balance). Finally, DellaVigna (2009) surveys field evidence for inattention to shipping fees, taxes, quality scores, and financial news.

# 3 Benchmark Model

This section develops the model structure used throughout the paper. The benchmark assumptions that are relaxed later are that (1) without regulation consumers are all sophisticated (inattentive but aware of their own inattention) and (2) all consumers have the same expected demand at the time of contracting. After describing the model, I derive optimal strategies of attentive and sophisticated inattentive consumers. Attentive consumers solve a dynamic programming problem and buy all units valued above a critical threshold which is a function of the date and past consumption. Sophisticated inattentive consumers cannot condition on past usage, so implement a usage independent threshold equal to expected marginal price. I define bill-shock regulation as a requirement for firms to disclose information that perfectly substitutes for attention, which in the context of the model is equivalent to requiring firms post transaction prices at the point of sale. Comparing equilibrium pricing with sophisticated inattentive consumers to that with attentive consumers thus illuminates the effect of bill-shock regulation. I show an equivalence result: neither sophisticated inattention nor bill-shock regulation affects substantive market outcomes.

## 3.1 Model

There are mass 1 of consumers and $N \geq 1$ firms. Each consumer privately learns a vector $\mathbf{x}$, describing his or her nonnegative idiosyncratic costs of doing business with each of the $N$ firms. At the contracting stage, $t = 0$, firms simultaneously offer contracts, and each consumer either signs a contract or receives her outside option (normalized to zero). At each later period, $t \in \{1, 2\}$, consumers privately learn a taste shock $v_t$ that measures a consumer's value for a unit of add-on service. Taste shocks $v_t$ are drawn independently with cumulative distribution $F$ that is atomless and has full support on $[0, 1]$. Then consumers (who have accepted a contract) make a binary quantity choice, $q_t \in \{0, 1\}$, by choosing whether or not to consume a unit of service. In the final

period, consumers contracted with firm $i$ make a payment $P\left(\mathbf{q}, \mathbf{p}^i\right)$ to firm $i$, as a function of past quantity choices $\mathbf{q} = (q_1, q_2)$. Firm $i$'s offered contract can be any deterministic price schedule.[8] A contract is characterized by a vector of prices $\mathbf{p}^i = \left(p_0^i, p_1^i, p_2^i, p_3^i\right)$ that includes a fixed fee $p_0^i$, base marginal charges $p_1^i$ and $p_2^i$ charged for purchasing a unit in either period 1 or 2 respectively, and an additional penalty fee $p_3^i$ charged if both units are purchased:

$$P(\mathbf{q}, \mathbf{p}^i) = p_0^i + p_1^i q_1 + p_2^i q_2 + p_3^i q_1 q_2. \tag{1}$$

Note that $p_3$ is a penalty fee if positive but a loyalty discount if negative.

A consumer's net utility is his gross utility less an idiosyncratic cost of doing business with the firm, such as a transportation cost. A consumer's gross utility $u$ from contracting with firm $i$ is a function of add-on quantity choices $q_t$, private taste shocks $v_t$, and payment to the firm:

$$u\left(\mathbf{q}, \mathbf{v}, \mathbf{p}^i\right) = q_1 v_1 + q_2 v_2 - P(\mathbf{q}, \mathbf{p}^i). \tag{2}$$

Conditional on contract prices $\mathbf{p}$, a consumer's optimal consumption strategy can be described by a function mapping valuations to quantity choices: $\mathbf{q}\left(\mathbf{v}; \mathbf{p}\right)$. A consumer's expected gross utility from contracting with firm $i$ and making optimal consumption choices thereafter is

$$U^i = E\left[u\left(\mathbf{q}\left(\mathbf{v}; \mathbf{p}^i\right), \mathbf{v}, \mathbf{p}^i\right)\right].$$

Similarly, let

$$S^i = E\left[\sum_{t=1}^{2} \left(v_t - c\right) q_t\left(\mathbf{v}; \mathbf{p}^i\right)\right]$$

be the expected gross social surplus (excluding transportation costs) generated by a consumer contracting with firm $i$ and making optimal consumption choices. (First-best expected gross surplus, $S^{FB} = 2\int_c^1 (v - c)\, f(v)\, dv$, is that generated by efficient consumption choices.)

A consumer's expected net utility, $U^i - x^i$, includes the consumer's idiosyncratic cost $x^i$ of doing business with firm $i$, which can be interpreted as a transportation cost or brand taste. Thus, fraction $G\left(U^i; U^{-i}\right)$ of consumers buy from firm $i$ if $i$ offers expected gross utility of $U^i$, while competitors offer $U^{-i}$:

$$G\left(U^i; U^{-i}\right) \equiv \Pr(U^i - x^i \geq \max_{j \neq i}\{U^j - x^j, 0\}).$$

The function $G\left(U^i; U^{-i}\right)$ describes the distribution of outside options among firm $i$'s potential customers, which is determined endogenously by offers of other firms $U^{-i}$. (I will often suppress dependence on $U^{-i}$ from the notation, and refer to $G\left(U^i\right)$ as firm $i$'s residual demand.)

[8] See Rochet and Stole (2002) for an insightful discussion of this assumption.

Firm profits per consumer equal payments less fixed costs (normalized to zero) and marginal cost $c \in (0,1)$ per unit served. Thus firm $i$'s expected profits are

$$\Pi^i = G\left(U^i; U^{-i}\right) E\left[P\left(\mathbf{q}\left(\mathbf{v}; \mathbf{p}^i\right), \mathbf{p}^i\right) - c\left(q_1\left(\mathbf{v}; \mathbf{p}^i\right) + q_2\left(\mathbf{v}; \mathbf{p}^i\right)\right)\right],$$

which can always be rewritten in terms of expected gross social surplus and consumer utility:

$$\Pi^i = G\left(U^i; U^{-i}\right)\left(S^i - U^i\right).$$

### 3.2 Consumer Strategies

The optimal consumption rule for an attentive consumer who signs a contract would be to consume a unit of service at time $t$ if and only if her value for the unit, $v_t$, exceeds a threshold that is a function of the date $t$ and past usage choices. Let the period one and two thresholds be $v_1^a$ and $v_2^a(q_1)$ respectively. Then, suppressing firm $i$ superscripts from prices, Proposition 1 describes an attentive consumer's optimal strategy.

**Proposition 1** *An attentive consumer's optimal strategy is to buy in period 2 if and only if $v_2$ exceeds the marginal price*

$$v_2^a(q_1) = p_2 + p_3 q_1, \tag{3}$$

*and to buy in period 1 if and only if $v_1$ exceeds the threshold*

$$v_1^a = p_1 + \int_{p_2}^{p_2+p_3} (1 - F(v))\, dv. \tag{4}$$

The intuition is that $v_1^a$ equals the expected marginal price conditional on purchase, $p_1 + (1 - F(p_2 + p_3))\, p_3$, plus the expected opportunity cost of foregone second-period purchases, $\int_{p_2}^{p_2+p_3} (v_2 - p_2)\, f(v_2)\, dv_2$.[9] Integrating by parts, this sum simplifies to equation (4).

An inattentive consumer cannot condition her strategy on past usage because she does not keep track of past usage. She exhibits imperfect recall, failing to remember $v_1$ or $q_1$ at time 2. Otherwise, sophisticated inattentive consumers are entirely rational and, in particular, are aware of their own inattention and plan accordingly when choosing a consumption strategy at time 0.[10]

[9] In contrast, consumers who exhibit *spotlighting* would myopically use a calling threshold equal to the transaction price of the current unit, $p_1$. Liebman and Zeckhauser (2004) analyze optimal nonlinear pricing given *spotlighting*. Aron-Dine, Einav, Finkelstein and Cullen (2012) provide evidence for partial spotlighting in healthcare consumption but reject complete spotlighting.

[10] In this paper, inattentive consumers are aware of prices when signing a contract, but are uncertain about marginal prices at the point of sale. Many models of add-on pricing examine the opposite situation, by assuming that consumers

A sophisticated inattentive consumer's behavior is the result of decision making by multiple selves.[11] The time-0 self signs a contract and plans a consumption strategy. The planned consumption strategy for time 2 is independent of $v_1$ and $q_1$ because the sophisticated time-0 self anticipates not recalling these variables at time 2. The time-1 self implements a consumption strategy at time 1, treating the planned time-2 consumption strategy as fixed.[12] Similarly, the time-2 self implements a consumption strategy at time 2, without recalling $v_1$ or $q_1$ from time 1 but knowing the planned strategy used at time 1. Proposition 2 describes a sophisticated inattentive consumer's optimal strategy. Importantly, this strategy is constrained time consistent (Battigalli, 1997, Proposition 3.5). The time-$t$ self will find it optimal to implement her previously planned strategy at time $t$, holding her strategy fixed at other times $t' \neq t$.

**Proposition 2** *A sophisticated inattentive consumer's optimal strategy is to buy in period t if and only if $v_t$ exceeds the expected marginal price of the current unit, $v_t^s$, that satisfies:*[13]

$$v_1^s = p_1 + (1 - F(v_2^s))\, p_3, \tag{5}$$

$$v_2^s = p_2 + (1 - F(v_1^s))\, p_3. \tag{6}$$

*If $|p_3| < 1/\bar{f}$, for $\bar{f} \equiv \max_v f(v)$, then equations (5)-(6) have a unique solution and hence are sufficient as well as necessary for $v_t^s$ to be optimal. Moreover, if $p_1 = p_2$, then that solution is a constant threshold $v_1^s = v_2^s = v^s$.*

The threshold strategy described by Proposition 2 is intuitive. It says that a consumer will buy a unit if and only if her value exceeds its expected marginal price, which is given by equations (5)-

---

are aware of transaction prices at the time of purchase, but are unaware of marginal prices or hidden fees at the time they make an ex ante decision to visit a store or purchase a base product (Diamond, 1971; Ellison, 2005; Gabaix and Laibson, 2006; Bubb and Kaufman, 2013; Heidhues, Kőszegi and Murooka, 2012, 2014).

[11] My approach is consistent with Piccione and Rubinstein's (1997) *modified multiselves* approach, or equivalently Aumann, Hart and Perry's (1997) *action optimality*. (Note that inattentive consumers exhibit imperfect recall but not Piccione and Rubinstein's (1997) *absentmindedness*.) Unlike multiself models of self control (Laibson, 1997) or addiction (Bernheim and Rangel, 2004), an inattentive consumer's multiple selves all have the same preferences. An alternate interpretation of the game is that the decision makers at times one and two are distinct family members who share a joint account but do not communicate purchases to each other between model periods 1 and 2. By assuming time-0 planning, I implicitly assume that they can communicate ex ante and coordinate on the best equilibrium. This assumption seems reasonable for joint account holders.

[12] If, at time 1 upon choosing $q_1 = 1$, a consumer were to try to revise her time 2 strategy to implement the threshold $v_2^a(1)$, this plan would be forgotten at time 2.

[13] Equations (5)-(6) are necessary up to the fact that all $v_t^s \geq 1$ are equivalent and all $v_t^s \leq 0$ are equivalent.

(6).[14] The first term in equations (5)-(6) is the base marginal charge $p_t$ that is applicable in period $t$. The second term captures the expected penalty fee or loyalty discount. Considering a purchase at time 1, the penalty fee $p_3$ applies if the consumer also purchases at time 2, which happens with probability $(1 - F(v_2^s))$ given the threshold strategy. Thus $(1 - F(v_2^s))\, p_3$ is the expected penalty fee or loyalty discount. Similarly, considering a purchase at time 2, the penalty fee applies if the consumer already purchased at time 1. An inattentive consumer cannot remember if she purchased last period, but given her own threshold strategy, she knows that the probability she made the purchase is $(1 - F(v_1^s))$. Thus, $(1 - F(v_1^s))\, p_3$ is the expected penalty fee or loyalty discount.

### 3.3 Bill-Shock Regulation

Suppose that a firm faced some inattentive consumers and had the option to disclose information at the point of sale that would be a perfect substitute for attention. In the context of the model this would mean disclosing, in period 2, whether or not the penalty fee applies.

**Definition 1** *Bill-shock regulation requires firms to disclose information at the point of sale that is a perfect substitute for attention.*

Within the model, which only incorporates two consumption opportunities, my definition of bill-shock regulation is equivalent to a *price-posting* requirement that firms disclose the transaction price applicable at the point of sale, which is similar to the requirement in the FCC's recent bill-shock agreement. Note that in a richer model with more than two purchase opportunities, a perfect substitute for attention would in general require reporting the full purchase history $\mathbf{q}^{t-1}$, which could be cumbersome relative to price-posting. However, in practice firms commonly set prices only as a function of total purchases $\sum_{t=1}^{T} q_t$ and, in this case, disclosing total purchases to date is sufficient to make inattentive consumers behave like attentive consumers. For instance, in the case of cellular phones, bill-shock regulation might require disclosures on the phone screen of the simple form "107 included minutes and 10 days remaining in billing cycle."[15]

[14] In contrast, consumers who exhibit *ironing* respond to the average price, for which Liebman and Zeckhauser (2004) analyze optimal nonlinear pricing. There is evidence for ironing in individual labor choices (Liebman and Zeckhauser, 2004) and electricity consumption (Ito, 2014). This may reflect the fact that many people do not realize that taxes and electricity prices are nonlinear, in which case average price is a good estimate of marginal price. In other settings, such as cellular phone service, debit card transactions, or health insurance, however, ironing is unlikely because consumers are fully aware that contract nonlinearities, such as an allowance of 'free' units or a deductible, exist.

[15] Cellular bills are typically only a function of total calling within each calling category (peak, off-peak, etc.), and do not depend on when during the billing cycle calls occurred. Note that I find that it can be optimal for firms to

An alternative regulation that could be considered would prohibit the use of penalty fees:

**Definition 2** *Banning penalty fees is the requirement that firms charge a constant marginal price as a function of usage:* $p_1 = p_2$ *and* $p_3 = 0$.

In both the benchmark model and the model of naive inattention in Section 4 it will be a result that bill-shock regulation leads to marginal cost pricing ($p_1 = p_2 = c$, $p_3 = 0$). In this case, the two forms of regulation have the same effect on market outcomes, as inattentive consumers (sophisticated or naive) behave as attentive consumers do when there are no penalty fees or loyalty discounts. Moreover, although the formal results in Section 5 are shown only for bill-shock regulation, the two regulations would have qualitatively similar effects in the price-discrimination model (see the working paper version of the paper (Grubb, 2012)).

### 3.4 Benchmark Result

Benchmark assumptions are that (1) without regulation all consumers are sophisticated (inattentive but aware of their own inattention) and (2) all consumers have the same expected demand at the time of contracting. In this setting, firms do best by setting marginal charges to implement the first-best allocation and extracting surplus through the fixed fee (balancing the trade-off between mark-up and volume in the standard way). This is made transparent by writing firm $i$'s profits in terms of expected gross social surplus and consumer utility: $\Pi^i = G\left(U^i; U^{-i}\right)\left(S^i - U^i\right)$. For any fixed utility offer $U^i$, firm profits are maximized by choosing marginal prices $p_1^i$, $p_2^i$, and $p_3^i$ to achieve first-best surplus, while adjusting the fixed fee $p_0^i$ to keep $U^i$ constant. This is true independent of regulation.

If consumers are attentive, achieving first-best allocations requires setting the marginal price of all units equal to marginal cost. If consumers are inattentive, however, achieving first-best allocations only requires that the expected marginal price equal marginal cost. As a result, inattention allows for contracts with surprise penalty fees or loyalty discounts in equilibrium but bill-shock regulation restricts equilibrium pricing and eliminates such penalty fees and loyalty discounts. Nevertheless, bill-shock regulation does not affect allocations or the division of surplus.

---

deviate from such simple pricing when consumers are attentive or naive. However, it is reasonable to believe that in practice firms are restricted to price as a function only of total usage because contract complexity is inherently expensive. Adding such a restriction to the model would not qualitatively change the main predictions about the consequences of regulation in Propositions 3 and 8 or Corollary 2. The only qualitative change in results about regulation is that bill-shock regulation could benefit naive consumers even in the absence of attentive consumers, contrary to Corollary 1.

**Proposition 3** *Assume that all consumers (1) are inattentive and sophisticated, and (2) draw values from the same distribution, $v_t \sim F(v_t)$. Without regulation, allocations are efficient and the set of equilibrium prices includes all three-part tariffs with $p_1 = p_2 = p$ and $p_3 = \frac{c-p}{1-F(c)}$ for all $p$ such that $|p - c| < (1 - F(c))/\bar{f}$. Bill-shock regulation or a penalty-fee ban would lead to marginal cost pricing ($p_1 = p_2 = c$ and $p_3 = 0$) but not affect allocations, firm profits, or consumer surplus.*

While Proposition 3 shows that surprise penalty fees or loyalty discounts can be weakly optimal if all consumers are inattentive, it also shows that firms cannot do strictly better than marginal cost pricing when consumers are sophisticated. This is because sophisticated consumers who are aware of their own inattention cannot be exploited. Thus, under benchmark assumptions, the firm cannot do better than maximizing total surplus. As a result, the predictions of Proposition 3 are hard to reconcile with firm behavior. In particular, Proposition 3 cannot explain why firms adopt surprise penalty fees over marginal cost pricing or explain firms' expressed aversion to bill-shock regulation (see footnote 4). Therefore, Proposition 3 indicates that the benchmark model is missing market features that are essential for understanding surprise penalty fees and bill-shock regulation. Hence, the rest of the paper is devoted to enriching the benchmark model.

### 3.5 Rational Inattention

As presented thus far, the model assumes consumer inattention exogenously. A natural extension is to allow consumers to endogenously choose whether to be attentive or inattentive by comparing costs and benefits. For this extension assume that, between periods one and two, each consumer can choose to invest costly effort $k$ to find out her past consumption $q_1$. (This may capture the time costs of calling customer service or logging into an account website.) This choice is made without knowledge of the past choice $q_1$ or the future realization $v_2$. Consumers who do invest $k$ to pay attention will recall first-period consumption in period two and correspond to attentive consumers. Consumers who choose not to invest $k$ correspond to inattentive consumers.[16] This model nests the preceding special case of exogenous inattention ($k = \infty$).

Sophisticated consumers correctly anticipate their own attention costs and plan whether or not to be attentive at time zero. Conditional on contract prices $\mathbf{p}$, let $\mathbf{q}^a(\mathbf{v}; \mathbf{p})$ and $\mathbf{q}^s(\mathbf{v}; \mathbf{p})$ be the consumption strategies of attentive and sophisticated inattentive consumers, described respectively

[16] I assume that consumers choose to be attentive or inattentive. In contrast, the macroeconomics literature on rational inattention begun by Sims (2003) typically models individuals who choose to be partially attentive and learn a noisy signal of the variable of interest (see Sims (2010) for an overview).

in Propositions 1 and 2. An attentive consumers earns expected gross utility of

$$U_a = E\left[u\left(\mathbf{q}^a\left(\mathbf{v};\mathbf{p}\right),\mathbf{v}\right)\right] = \begin{array}{c} -p_0 + \int_{v_1^a}^1 (v-p_1) f(v)\, dv + F\left(v_1^a\right) \int_{p_2}^1 (v-p_2) f(v)\, dv \\ + \left(1-F\left(v_1^a\right)\right) \int_{p_2+p_3}^1 (v-p_2-p_3) f(v)\, dv \end{array} . \quad (7)$$

Similarly, a sophisticated consumer who chooses to be inattentive earns expected gross utility of

$$U_s = E\left[u\left(\mathbf{q}^s\left(\mathbf{v};\mathbf{p}\right),\mathbf{v}\right)\right] = -p_0 + \sum_{t\in\{1,2\}} \int_{v_t^s}^1 (v-p_t)\, dF(v) - p_3\left(1-F\left(v_1^s\right)\right)\left(1-F\left(v_2^s\right)\right). \quad (8)$$

When choosing whether or not to be attentive, sophisticated consumers compare the cost of attention, $k$, with the benefit of attention $\Delta_s = U_a - U_s$. Consumers choose to be inattentive if $\Delta_s \leq k$. (I assume that consumers choose to be inattentive when indifferent for technical convenience.)

Proposition 3 easily extends to the case of rational inattention for any $k > 0$, with two changes: (1) First, inattention would be a result rather than an assumption. In equilibrium, firms would offer prices that induce sophisticated consumers to choose to be inattentive. This would occur because the attention cost is socially wasteful. (2) Second, the restriction on $p$ for $p_3 = (c-p)/(1-F(c))$ to be a feasible equilibrium penalty fee would depend on $k$ as well as $\bar{f}$.[17] In the rest of the paper I assume rational inattention where it substantially affects my results (Section 4) but exogenously impose inattention for simplicity where it does not (Section 5).

# 4 Naive Consumers

In this section, I first characterize behavior by consumers who naively fail to anticipate their own inattention. Surprise penalty fees lead such naively inattentive consumers to underestimate the likelihood of paying penalty fees (via overconfidence). Symmetrically, loyalty discounts (or negative penalty fees) lead customers to overestimate the probability of receiving a discount (via underconfidence). Next, I develop the first and second main results concerning equilibrium pricing and bill-shock regulation. The first main result is that firms find it strictly optimal to charge surprise penalty fees or loyalty discounts in order to exploit consumer naivete. Bill-shock regulation restores marginal cost pricing but, when all consumers are naively inattentive, it does not affect welfare in a Hotelling duopoly. When attentive and naively inattentive consumers are in the same market, however, the second main result shows that welfare implications for bill-shock regulation differ

[17] Notice that for $p_3 = 0$, there is no benefit to attention ($\Delta_s = 0$) and consumers choose to be inattentive. Moreover, holding $v_t^s$ constant, $\Delta_s$ varies continuously in $p_3$. Thus for any consumption thresholds $v_1^s$ and $v_2^s$ that a firm desires to implement (by choosing $p_1$ and $p_2$ to satisfy equations (5)-(6) and $|p_3| < 1/\bar{f}$), there exist bounds $\underline{p}\left(v_1^s, v_2^s\right) < 0$ and $\bar{p}\left(v_1^s, v_2^s\right) > 0$ such that for any $p_3 \in \left[\underline{p}, \bar{p}\right]$, sophisticated consumers choose to be inattentive. Thus $p$ would have satisfy $|c-p|/(1-F(c)) \in \left[\underline{p}(c,c), \bar{p}(c,c)\right]$ to ensure $p_3 = (c-p)/(1-F(c))$ is a feasible equilibrium penalty fee.

substantially. Bill-shock regulation both ends cross subsidies from naively inattentive to attentive consumers and raises average consumer surplus by eliminating allocative distortions that arise when firms try to cater to both groups. Moreover, if attentive consumers are more price sensitive than the naively inattentive then bill-shock regulation can benefit all consumers by stiffening competition.

### 4.1 Naive Consumer Behavior

Until now, I have assumed that inattentive consumers are sophisticated and correctly anticipate their own inattention. In this section, however, I assume that consumers are naive about their own inattention by adapting the model of rational inattention outlined in Section 3.5. Naivete arises because consumers underestimate their own costs of attention. In particular, while their true cost of attention, $k$, is strictly positive, naive consumers initially believe it to be zero and hence always anticipate being attentive. As a result, in periods 0 and 1 when consumers choose a contract and an initial consumption threshold, naive consumers believe that in period 2 they will recall their period 1 consumption. Between periods 1 and 2, however, naive consumers realize the true cost of attention and will choose to be inattentive if it is too high. In this case, contrary to their own expectations, naive consumers will not recall past consumption in period 2.

Given the assumption that naive consumers expect their attention costs to be zero, it follows that they expect to follow the attentive consumption rule described by Proposition 1. In period 1, they do so, choosing the same initial consumption threshold $v_1^a$ as attentive consumers, described in equation (4). Between periods 1 and 2, they realize the true cost of attention is $k > 0$ and reevaluate the decision to be attentive. Following through on the plan to be attentive yields gross expected utility $U_a$ (equation (7)) less the cost of attention $k$. Changing course and being inattentive means implementing an alternative consumption strategy $\mathbf{q}^n(\mathbf{v};\mathbf{p})$ with a second period consumption threshold $v_2^n$ that is independent of past consumption. The optimal threshold $v_2^n$ equals the expected marginal price conditional on the initial consumption threshold $v_1^a$. This yields expected gross utility of $U_n = E\left[u\left(\mathbf{q}^n(\mathbf{v};\mathbf{p}),\mathbf{v}\right)\right]$,

$$U_n = -p_0 + \int_{v_1^a}^{1}(v-p_1)\,dF(v) + \int_{v_2^n}^{1}(v-p_2)\,dF(v) - p_3\left(1-F\left(v_1^a\right)\right)\left(1-F\left(v_2^n\right)\right). \tag{9}$$

As a result, while naive consumers always anticipate being attentive, they will instead choose to be inattentive whenever $\Delta_n = U_a - U_n \leq k$. This is summarized by Proposition 4.

**Proposition 4** *A naive consumer anticipates being attentive and buying in period $t$ if and only if $v_t$ exceeds the attentive threshold $v_t^a$ described by Proposition 1. If $\Delta_n > k$, this expectation is correct. If $\Delta_n \leq k$, however, a naive consumer is inattentive: In period 1, the naive consumer*

*buys if $v_1$ exceeds the attentive threshold $v_1^a$ given by equation (4). In period 2, however, the naive consumer buys if $v_2$ exceeds the threshold*

$$v_2^n = p_2 + p_3 \left(1 - F\left(v_1^a\right)\right). \tag{10}$$

Consumer naivete leads to a difference between perceived and true expected utilities at time zero. Naive consumers always perceive their expected gross utility to be $U_a$. The true expected gross utility is instead $\max\{U_a - k, U_n\}$, meaning that naive consumers overestimate contract value by $\min\{\Delta_n, k\}$. If $\Delta_n \leq k$, so that naive consumers choose to be inattentive, then consumers overvalue an offered contract by $\Delta_n$:

$$\Delta_n = F\left(v_1^a\right) \int_{p_2}^{v_2^n} \left(v - p_2\right) f\left(v\right) dv + \left(1 - F\left(v_1^a\right)\right) \int_{v_2^n}^{p_2+p_3} \left(p_2 + p_3 - v\right) f\left(v\right) dv. \tag{11}$$

Naively inattentive consumers' over-valuation of a contract stems from the difference between their true consumption behavior, $\mathbf{q}^n\left(\mathbf{v};\mathbf{p}\right)$, and their anticipated consumption behavior, $\mathbf{q}^a\left(\mathbf{v};\mathbf{p}\right)$. When $p_3$ is zero, the two consumption rules coincide and there is no overvaluation ($\Delta_n = 0$). When $p_3$ differs from zero (either positive or negative) the true and anticipated consumption rules diverge and there is overvaluation.

To understand the role of $p_3$ on consumer overvaluation, consider the penalty fee ($p_3 > 0$) and loyalty discount ($p_3 < 0$) cases separately. First, consider a penalty fee. Penalty fees lead naively inattentive consumers to make two mistakes that both lead to contract overvaluation. First, they underestimate the probability of buying a unit in period two when it is the second unit that triggers a penalty fee (when $q_1 = 1$). Second, they overestimate the probability of buying a unit in period two when it is an inexpensive first unit (when $q_1 = 0$).

The first mistake means underestimating the probability of consuming two units and paying a penalty fee. In particular, a naively inattentive consumer anticipates paying a penalty fee only with probability $\left(1 - F\left(v_1^a\right)\right)\left(1 - F\left(p_2 + p_3\right)\right)$ but ends up paying it with the larger probability $\left(1 - F\left(v_1^a\right)\right)\left(1 - F\left(v_2^n\right)\right)$.[18] This follows from the fact that a naively inattentive consumer believes she will pay attention, recognize when the penalty fee applies, and be especially selective about period 2 consumption when it does. However, she does not pay attention and sometimes pays the penalty without realizing it in cases when her value is below the marginal price: $v_2 < p_2 + p_3$.

The second mistake arises for similar reasons. A naively inattentive consumer believes she will pay attention, recognize when the penalty fee does not apply in period two, and consume more aggressively when it does not. In fact, however, she does not pay attention and sometimes fails to make a purchase in period two even though her value exceeds the marginal price: $v_2 > p_2$.

[18]The latter probability is larger because $v_2^n < p_2 + p_3$ whenever $p_3 > 0$ and $v_1^a > 0$ by equation (10).

Together, both mistakes imply that naively inattentive consumers overestimate the probability of buying exactly one unit when facing a contract with a penalty fee. As a result, penalty fees lead naively inattentive consumers to endogenously exhibit overconfidence in the sense that they overestimate the precision of their forecasts of total consumption ($Q = q_1 + q_2$) and underestimate the variability in their total consumption. (Formally, overconfidence corresponds to the fact that $H^*(Q)$ crosses $H(Q)$ once from below, where $H(Q)$ and $H^*(Q)$ are respectively the true and perceived cumulative distribution functions of total consumption (Grubb, 2009).) Thus naive inattention may explain the presence of overconfidence that Grubb and Osborne (forthcoming) estimate from cellular phone billing data.[19]

Next, consider the case of a loyalty discount ($p_3 < 0$). In this case, naively inattentive consumers make exactly the opposite mistakes as they would with a penalty fee. First, naively inattentive consumers overestimate the probability of consuming both units and receiving a loyalty discount. In particular, a naively inattentive consumer anticipates receiving a loyalty discount with probability $(1 - F(v_1^a))(1 - F(p_2 + p_3))$ but ends up receiving it with the smaller probability $(1 - F(v_1^a))(1 - F(v_2^n))$.[20] This follows from the fact that a naively inattentive consumer believes she will pay attention, recognize when a loyalty discount is available, and be more aggressive about period 2 consumption when it is. However, she does not pay attention and sometimes misses the loyalty discount without realizing it in cases where her value is above the marginal price: $v_2 > p_2 + p_3$. Second, for similar reasons, naively inattentive consumers overestimate the probability of avoiding the high cost of a first unit by consuming zero units. Together, these mistakes imply that loyalty discounts lead naively inattentive consumers to overvalue contracts and exhibit underconfidence by overestimating the variability of their total consumption. (Formally $H^*(Q)$ crosses $H(Q)$ once from above.)

The preceding discussion is formalized in the following proposition.

**Proposition 5** *Consider a contract for which* $\{v_1^a, v_2^n\} \in (0,1)^2$ *and* $\Delta_n \leq k$. *Naive consumers are inattentive. If there is a penalty fee* ($p_3 > 0$) *then naive consumers are overconfident, overvalue the contract, and underestimate the probability of paying penalty fees. If there is a loyalty discount*

[19] Experimental evidence shows that individuals are overconfident about the precision of their own predictions even in situations in which naive inattention could not be the cause (Lichtenstein, Fischhoff and Phillips, 1982). Existing behavioral IO models incorporate incorrect expectations about future consumption due to a variety of causes other than inattention, including exogenously biased beliefs (Eliaz and Spiegler, 2006, 2008; Grubb, 2009), myopia (Gabaix and Laibson, 2006; Miao, 2010), naive quasi-hyperbolic-discounting (DellaVigna and Malmendier, 2004), or naivete about sales advice (Inderst and Ottaviani, 2013). See Spiegler (2011) for a survey.

[20] The latter probability is smaller because $v_2^n > p_2 + p_3$ whenever $p_3 < 0$ and $v_1^a > 0$ by equation (10).

*($p_3 < 0$) then naive consumers are underconfident, overvalue the contract, and overestimate the probability of receiving a loyalty discount.*

### 4.2 Pricing without bill-shock regulation (naively inattentive case)

The first pricing result is that firms always find it optimal to set penalty fees or loyalty discounts sufficiently close to zero to induce naive consumers to be inattentive. In particular, any contract with $p_3 \neq 0$ that induces consumers to be attentive would be strictly less profitable than marginal cost pricing. This is implied by Proposition 3, which shows that marginal cost pricing is uniquely optimal under bill-shock regulation, or equivalently when consumers are attentive. Thus a firm's profit maximization problem can be written with the constraint $\Delta_n \leq k$ ensuring that naive consumers are inattentive.

To derive a firm's best response, it is useful to reframe the firm's problem in two ways. First, think of the firm choosing perceived utility level $U_a$, so that the fixed fee $p_0$ is given by equation (7). The fraction of naive consumers who buy from the firm, $G_n(U_a)$, depends on $U_a$ because naive consumers anticipate being attentive at no cost, following consumption rule $q^a$, and receiving expected gross utility $U_a$. Second, think of the firm choosing naively inattentive consumers' first and second period consumption thresholds $v_1^a$ and $v_2^n$ and then setting marginal prices $p_1$ and $p_2$ to implement them. Rewriting equations (4) and (10) this implies $p_1$ and $p_2$ satisfy:

$$p_1 = v_1^a - \int_{p_2}^{p_2+p_3} (1 - F(v))\, dv, \tag{12}$$

$$p_2 = v_2^n - p_3 (1 - F(v_1^a)). \tag{13}$$

The firm's markup, $S_n - U_n$, depends on the true expected gross utility $U_n$ (equation (9)) and the expected gross social surplus

$$S_n = \int_{v_1^a}^{1} (v - c)\, dF(v) + \int_{v_2^n}^{1} (v - c)\, dF(v), \tag{14}$$

generated by naively inattentive consumers. Thus a firm's expected profits are $\Pi = G_n(U_a)(S_n - U_n)$. Substituting $U_n = U_a - \Delta_n$, the firm's problem can be written as

$$\max_{\substack{U_a, v_1^a, v_2^n, p_3 \\ \text{such that } \Delta_n \leq k}} G_n(U_a)\left(S_n(v_1^a, v_2^n) - U_a + \Delta_n(v_1^a, v_2^n, p_3)\right), \tag{15}$$

where $p_0$, $\Delta_n$, $p_1$, $p_2$, and $S_n$ satisfy equations (7) and (11)-(14).

Inspection of the firm's problem shows that maximizing profits entails choosing $v_1^a$, $v_2^n$, and $p_3$ to maximize the sum of surplus $S_n$ and the amount $\Delta_n$ by which consumers overvalue the offered contract subject to $\Delta_n \leq k$. Importantly, the penalty fee $p_3$ only enters the firm's problem through

consumers' overvaluation $\Delta_n$, and is thus chosen solely to maximize consumers' overvaluation of the contract.

As $p_3$ becomes increasingly different from zero (either positive or negative) the consumers' true and anticipated consumption rules diverge and contract overvaluation increases. This follows from the derivative of $\Delta_n$ with respect to $p_3$,

$$\frac{d\Delta_n}{dp_3} = F\left(v_1^a\right)\left(1 - F\left(v_1^a\right)\right)\left(F\left(p_2 + p_3\right) - F\left(p_2\right)\right), \tag{16}$$

which is negative for $p_3 < 0$ and positive for $p_3 > 0$.[21] Importantly, for sufficiently large $|p_3|$, the term $\left(F\left(p_2 + p_3\right) - F\left(p_2\right)\right)$ in equation (16) equals one. Beyond this point, equation (16) shows that $\Delta_n$ increases linearly in $|p_3|$. Thus the firm can always choose a loyalty discount or penalty fee to satisfy the constraint $\Delta_n \leq k$ with equality and it will be optimal to do so. This insight can be used to simplify the firm's problem.

Define the critical loyalty discount that achieves $\Delta_n = k$ as $p^{\min}$ and the corresponding critical penalty fee as $p^{\max}$:

$$\begin{aligned} p^{\min}\left(v_1^a, v_2^n\right) &\equiv \left\{p_3 : \Delta_n\left(v_1^a, v_2^n, p_3\right) = k \text{ and } p_3 < 0\right\}, \\ p^{\max}\left(v_1^a, v_2^n\right) &\equiv \left\{p_3 : \Delta_n\left(v_1^a, v_2^n, p_3\right) = k \text{ and } p_3 > 0\right\}. \end{aligned}$$

Then the firm's profit maximization problem reduces to

$$\max_{U_a, v_1^a, v_2^n} G_n\left(U_a\right)\left(S_n\left(v_1^a, v_2^n\right) - U_a + k\right), \tag{17}$$

where $p_3 \in \left\{p^{\min}\left(v_1^a, v_2^n\right), p^{\max}\left(v_1^a, v_2^n\right)\right\}$ and $p_0$, $p_1$, $p_2$, and $S_n$ satisfy equations (7) and (12)-(14). By inspection of the firm's objective in equation (17) it is clear that profits are maximized by setting consumption thresholds equal to marginal cost to achieve first best surplus: $v_1^a = v_2^n = c$.

Proposition 6 summarizes the preceding results.

**Proposition 6** *Firm Best Response: Given exogenous residual demand $G_n\left(U_a\right)$, optimal contracts satisfy the following: (1) Consumption thresholds are efficient: $v_1^a = v_2^n = c$. (2) Firms are indifferent between offering the maximum feasible surprise loyalty discount, $p_3 = p^{\min}\left(c, c\right) < 0$, or charging the maximum feasible surprise penalty fee $p_3 = p^{\max}\left(c, c\right) > 0$. In either case, the firm finds it strictly optimal not to disclose the loyalty discount or penalty fee at the point of sale.*

Proposition 6 includes part of the first main result: It suggests that consumer naivete about inattention could explain both the existence of loyalty discount programs and the existence of

[21] Proposition 6 shows that, at equilibrium prices, $v_1^a = v_2^n = c \in (0, 1)$ and therefore $F\left(v_1^a\right)\left(1 - F\left(v_1^a\right)\right) > 0$ and $F\left(p_2 + p_3\right) \neq F\left(p_2\right)$ if $p_3 \neq 0$.

penalty fees in cases for which the loyalty discounts and penalties are not disclosed at the point of sale.[22] While the result provides no indication of why firms might favor one over the other, the following proposition (illustrated by Example 1 at the end of Section 4) illuminates an important difference between the two pricing structures.

**Proposition 7** *If the market is sufficiently competitive (by which I mean if equilibrium utility offers $U_a$ are sufficiently close to their perfectly competitive level $S^{FB}$) then equilibrium penalty-fee contracts have positive fixed fees but equilibrium loyalty-discount contracts have negative fixed fees.*

Charging negative fixed fees is likely to be costly as there is always likely to be a pool of customers with no value for the service who would collect the fixed-fee subsidy without making further purchases. This could explain why firms often choose penalty fees over loyalty discounts.

### 4.3 Consequences of bill-shock regulation

Proposition 6 shows that, without regulation, inattention and naivete lead firms to create surprise penalty fees or surprise loyalty discounts but do not distort consumption thresholds away from first best. In contrast, bill-shock regulation makes naively inattentive consumers indistinguishable from attentive consumers. In particular, as bill-shock alerts provide a perfect substitute for attention (by assumption), consumer naivete has no effect. Thus (following the benchmark results in Proposition 3) bill-shock regulation leads to marginal cost pricing and first-best surplus under any market structure. By comparing the two cases, Corollary 1 shows no clear benefit from bill-shock regulation despite consumer naivete. To state the Corollary, I first define a Hotelling duopoly.

**Definition 3** *In a Hotelling duopoly, a firm is on each end of a uniform Hotelling line. Linear transport costs $\tau$ are sufficiently small for strict full-market-coverage with or without regulation.*[23]

Next, Corollary 1 completes the first main result:

[22] While naive inattention is a novel explanation for loyalty discounts, it relates to the more straight-forward prediction that mail-in rebates should be offered to consumers who underestimate the likelihood of redeeming them (Jolson, Wiener and Rosecky, 1987; Soman, 1998; Khouja, 2006). Other explanations for loyalty discounts are that they can induce one-stop shopping (Banerjee and Summers, 1987; Cairns and Galbraith, 1990) and can be profitable if fixed fees cannot be charged (Crémer, 1984; Bulkley, 1992; Caminal and Claici, 2007).

[23] *Strict* full-market-coverage requires that every consumer *strictly* prefer the best offer to her outside option. Results generalize to oligopolies with more than two firms. The important simplifying assumption is full market coverage, which implies that regulation affects allocations on the intensive margin but not the extensive margin.

**Corollary 1** *Assume attention costs are $k > 0$ but consumers naively believe attention is costless. (1) Bill-shock regulation eliminates surprise penalty fees and loyalty discounts. (2) In a Hotelling duopoly, firm profits, consumer surplus, and social welfare are unchanged by bill-shock regulation.*

The fact that Corollary 1 predicts no welfare consequences of bill-shock regulation depends importantly on the assumption that transportation costs are sufficiently small for full market coverage with or without regulation ($\tau < (2/3)S^{FB}$). This implies that industry profits are $\tau$ both before and after regulation. Thus, absent regulation, competition protects consumers from exploitation despite their naivete and overvaluation of contracts. Regulation has no additional benefit to consumers; it simply shifts the source of firm revenue from penalty fees to fixed fees. (This matches claims made by some critics of bill-shock regulation (Federal Reserve Board, 2009).)[24] In industries with more market power, eliminating consumers' contract overvaluation could reduce market coverage, increasing consumer welfare and lowering firm profits.[25] More importantly, Corollary 1 also depends heavily on assuming all consumers are naively inattentive. The following section relaxes this assumption.

### 4.4 Heterogenous levels of attention

To develop the second main result, suppose now that there are two types of consumers in the market: Let there be $\alpha_a > 0$ attentive consumers with zero cost of attention and $\alpha_n = 1 - \alpha_a > 0$ naive consumers with attention cost $k > 0$ who mistakenly believe attention is costless. Attentive and naive consumers will always choose the same contract and thus the firm continues to offer a single contract. As when all consumers are naive, it is optimal for the firm to induce naive consumers

[24]Jamie Dimon, CEO of JPMorgan Chase said, "If you're a restaurant and you can't charge for the soda, you're going to charge more for the burger. Over time, it will all be repriced into the business" (Dash and Schwartz, 2010). Jamie Dimon's logic and the supporting result in Corollary 1 both implicitly rely on the assumption that penalty fee profits can be competed away through reduced fixed fees. Farrell and Klemperer (2007), Miao (2010), Heidhues et al. (2012, 2014), and Armstrong and Vickers (2012) show that profits from aftermarket sales are not necessarily competed away in primary market competition if the danger of attracting worthless customers prevents firms charging negative prices for primary goods. In this spirit, in the working paper version of this paper (Grubb, 2012), I show that bill-shock regulation can stiffen price competition to inattentive consumers' benefit when total prices must be nonnegative.

[25]The extreme case is that of a socially wasteful product with $c \geq 1$ and $\tau > S^{FB} = 0$. For $k > 0$ sufficiently large, the market would operate for such a product in the absence of regulation, inducing value destroying consumption for the sole purpose of exploiting consumer naivete. Bill-shock regulation would lead to the socially efficient shut-down of such a market. The failure of competition to protect behavioral consumers from exploitation when products are socially wasteful has also been noted by Heidhues et al. (2012, 2014).

to be inattentive by limiting $\Delta_n \leq k$. Moreover, penalty fees and loyalty discounts remain useful for exploiting naivete. Now however, setting $p_3 \neq 0$ also bears a cost: while charging $p_3 \neq 0$ can raise the profits earned from naively inattentive consumers it limits profits earned from attentive consumers. (The loss in profits earned from attentive consumers follows because charging $p_3 \neq 0$ necessarily distorts an attentive consumer's consumption choices in period 2, thereby reducing surplus and profits.) A firm's revised objective is therefore

$$\max_{\substack{U_a, v_1^a, v_2^n, p_3 \\ \text{such that } \Delta_n \leq k}} \left( \begin{array}{c} \alpha_a G_a(U_a)\left(S_a\left(v_1^a, v_2^n, p_3\right) - U_a\right) \\ +\alpha_n G_n(U_a)\left(S_n\left(v_1^a, v_2^n\right) - U_a + \Delta_n\left(v_1^a, v_2^n, p_3\right)\right) \end{array} \right), \tag{18}$$

where $G_a(U_a)$ and $G_n(U_a)$ are the respective shares of attentive and naive consumers attracted by utility offer $U_a$, expected gross social surplus from attentive consumers is

$$S_a\left(v_1^a, v_2^n\right) = \int_{v_1^a}^{1} (v-c)\, dF(v) + F\left(v_1^a\right) \int_{p_2}^{1} (v-c)\, dF(v) + \left(1 - F\left(v_1^a\right)\right) \int_{p_2+p_3}^{1} (v-c)\, dF(v), \tag{19}$$

and, as before, $p_0$, $\Delta_n$, $p_1$, $p_2$, and $S_n$ satisfy equations (7) and (11)-(14).

The costs and benefits of setting $p_3 \neq 0$ are proportional to the fractions of attentive and naive consumers, respectively. Thus, if there are sufficiently few naive consumers, it will be optimal not to use penalty fees or loyalty discounts ($p_3 = 0$). However, if there are more naive consumers than attentive consumers ($\alpha_n > \alpha_a$) then a penalty fee or loyalty discount is always optimal ($p_3 \neq 0$). Moreover, although naive consumers may consume efficiently in equilibrium ($v_1^a = v_2^n = c$),[26] consumption by attentive consumers is always inefficient in period 2 whenever $p_3 \neq 0$. As a result, bill-shock regulation always increases social welfare.

**Proposition 8** *Consider a Hotelling duopoly in which naive consumers have attention cost $k > 0$ and there are more naive than attentive consumers ($\alpha_n > \alpha_a > 0$). (1) Without regulation, at least one firm offers a surprise loyalty discount ($p_3 < 0$) or charges a surprise penalty fee ($p_3 > 0$). In either case, naive consumers are inattentive. (2) Bill-shock regulation strictly raises social surplus. (3) If naive and attentive consumers have equal transportation costs ($\tau_a = \tau_n = \tau$) then the unique equilibrium is symmetric and bill-shock regulation benefits naive consumers, harms attentive consumers, and does not affect firm profits.*

In the special case of equal transportation costs, Proposition 8 shows that although average consumer surplus increases, attentive consumers are made worse off. This follows because bill-shock

[26] This is no longer always true given the presence of attentive consumers.

regulation stops naive consumers from cross-subsidizing attentive consumers.[27] (As all consumers expect to be attentive ex ante, this also means that regulation might have little consumer support.)

An interesting alternative to consider is the case in which attentive consumers have lower transportation costs than the naively inattentive ($\tau_a < \tau_n$). This is a reasonable assumption if attentive consumers are attentive precisely because they have a low cost of time and can therefore more easily devote time both to keeping track of past purchases and shopping for a good price. In this case two new possibilities arise. (1) First, bill-shock regulation may lower industry profits and actually benefit attentive consumers by stiffening competition. (2) Second, without regulation, asymmetric equilibria may arise in which one firm exploits naive consumers with surprise penalty fees or loyalty discounts while the other offers fair pricing at marginal cost. Both possibilities are illustrated by Example 1.

The fact that bill-shock regulation may lower industry profits if $\tau_a < \tau_n$ may explain why industry trade groups lobby against regulation. For intuition, notice that when attentive consumers are more price sensitive ($\tau_a < \tau_n$) firms face adverse selection when lowering prices. Any price cut will attract disproportionately more attentive consumers than naive consumers. As attentive consumers are less profitable than naive consumers this softens firms' incentives to compete on price and leads to high markups in equilibrium.[28] Bill-shock regulation eliminates adverse selection and stiffens competition by making attentive and naive consumers equally profitable. By demonstrating this possibility, Example 1 completes the second main result:

**Example 1** *Values are uniformly distributed on* $[0,1]$*, marginal cost is* $c = 1/2$*, equal numbers of consumers are attentive and naive* ($\alpha_a = \alpha_n = 1/2$)*, naive consumers have attention cost* $k = 9/128 \approx 0.07$*, and there is a Hotelling duopoly with transport costs* $(\tau_a, \tau_n) = (1/16, 1/8)$*.*

In this example, allocations are not distorted for naive consumers ($v_1^a = v_2^n = c$) and distortions arise only for attentive consumers in period 2. (See Appendix B.10 regarding the derivation of the solution.) For $v_1^a = v_2^n = c = 1/2$ and $k = 9/128$, the maximum penalty fee or loyalty discount is $|p_3| = 3/4$. If more than half of a firm's customers are attentive then marginal cost pricing is optimal. If less than half of a firm's customers are attentive then it is optimal to charge either the maximum penalty fee or to offer the maximum loyalty discount.

Without regulation, both pure strategy equilibria are asymmetric and identical up to a relabeling of firms. Equilibrium outcomes are summarized in Table 1. Firm A offers marginal cost pricing

[27] Cross-subsidization also occurs in other add-on pricing models with naive consumers (Gabaix and Laibson, 2006; Bubb and Kaufman, 2013; Armstrong and Vickers, 2012).

[28] This mechanism is similar to that in Ellison (2005).

Table 1: Equilibrium Outcomes in Example 1

| | No Regulation | | Bill-Shock Regulation |
|---|---|---|---|
| | Firm 1 | Firm 2 | Firm 1/2 |
| Penalty fee or Loyalty discount | no | yes | no |
| $\begin{bmatrix} p_0 \\ p_1 \\ p_3 \end{bmatrix}$ (note $p_1 = p_2$) | $\begin{bmatrix} 0.091 \\ 1/2 \\ 0 \end{bmatrix}$ | $\begin{bmatrix} 0.36 \\ 1/8 \\ 3/4 \end{bmatrix}$ or $\begin{bmatrix} -0.02 \\ 7/8 \\ -3/4 \end{bmatrix}$ | $\begin{bmatrix} 0.083 \\ 1/2 \\ 0 \end{bmatrix}$ |
| Gross Utility: $U_a$, $U_n$ | 0.159, 0.159 | 0.151, 0.081 | 0.167, 0.167 |
| Market Shares (a,n) | 56%, 53% | 44%, 47% | 50%, 50% |
| Gross Surplus: $S_a$, $S_n$ | 0.25, 0.25 | 0.18, 0.25 | 0.25, 0.25 |
| Markups: $\mu_a$, $\mu_n$ | 0.09, 0.09 | 0.03, 0.17 | 0.08, 0.08 |
| Pct. own consumers naive | 48% | 52% | 0% |
| Firm Profits | 0.050 | 0.046 | 0.042 |
| Total Profits; Consumer Surplus | 0.096; 0.115 | | 0.083; 0.143 |
| Net Social Surplus | 0.211 | | 0.227 |

with $U_a^A = U_n^A = 0.159$ and $S_a^A = S_n^A = 1/4$ and earns $\Pi^A = 0.050$. Firm B offers the maximum loyalty discount or the maximum penalty ($|p_3| = 3/4$) with $U_a^B = 0.151$, $U_n^B = 0.081$, $S_a^B = 0.18$, and $S_n^B = 1/4$ and earns $\Pi^B = 0.046$. At these prices, Firm A earns a markup of 0.09 on both types whereas Firm B earns a markup of 0.17 on naive consumers but of 0.03 on attentive consumers. Thus Firm B faces adverse selection when it competes on the fixed fee while Firm A does not, which explains why Firm B is a softer competitor and offers $U_a^B < U_a^A$ in equilibrium. As a result, Firm A has larger market share than B, particularly among attentive consumers. Firm A attracts 56% of attentive consumers and 53% of naive consumers. Thus Firm A's customers are predominantly attentive, and hence marginal cost pricing is optimal for A. At the same time, Firm B's customers are predominantly naive and hence $|p_3| = 3/4$ is optimal for B.

Given bill-shock regulation, equilibrium is symmetric with both firms offering marginal-cost pricing, splitting the market, and earning profits of 0.042 each. For both types, gross expected surplus is $S_a = S_a = 1/4$ and gross expected utility is $U_a = U_n = 0.167$. Thus, bill-shock regulation increases social surplus by 0.016, reduces industry profits by 0.012, raises average consumer surplus by 0.028, and benefits all consumers.

A final interesting aspect of Example 1 concerns fixed fees. In the unregulated equilibrium, Firm B is indifferent between a loyalty discount $(p_0, p_1, p_2, p_3) = (-0.02, 7/8, 7/8, -3/4)$ and a penalty fee $(p_0, p_1, p_2, p_3) = (0.36, 1/8, 1/8, 3/4)$. However, the loyalty discount requires a negative fixed fee while the penalty fee does not. This feature of the example is an illustration of Proposition 7. As

a result, the penalty fee contract may be better as it precludes arbitrage opportunities.

# 5 Price Discrimination via Multiple Contracts

In this section, I return to the benchmark assumption that all consumers are sophisticated and for simplicity impose inattention exogenously ($k = \infty$).[29] However, I relax the assumption of homogeneous demand imposed in the benchmark model and show that surprise penalty fees help firms price discriminate between consumers with ex ante low and high demand. In this alternative setting, I show that bill-shock regulation is socially harmful in fairly-competitive markets. This contrasts sharply to the preceding setting with naively inattentive consumers in which bill-shock regulation was always weakly beneficial. To analyze the effect of bill-shock regulation, I analyze equilibria with and without the regulation and compare them. I begin by characterizing equilibrium with bill-shock regulation, which is simpler and corresponds to all consumers being attentive.

## 5.1 Model

This section extends the benchmark model to include low and high demand consumers. Prior to choosing a contract, each consumer privately receives one of two private signals $\theta \in \{L, H\}$, where $\Pr(\theta = H) = \beta$. As a result, each firm $i$ simultaneously offers a menu with a choice of two contracts. Each contract is characterized by a vector of prices $\mathbf{p}^i_\theta = \left(p^i_{0\theta}, p^i_{1\theta}, p^i_{2\theta}, p^i_{3\theta}\right)$, which specifies payments as in equation (1). As before, at each later period $t \in \{1, 2\}$, a consumer privately learns her value $v_t$ for a unit of add-on service. Conditional on receiving signal $\theta$, a consumer's values $v_t$ are drawn independently with cumulative conditional distribution $F_\theta$, which is atomless and has full support on $[0, 1]$. Distributions $F_L$ and $F_H$ are ranked by first-order stochastic dominance, $F_L(v) \geq F_H(v)$, and the ranking is strict at marginal cost, $F_L(c) > F_H(c)$, where $c \in (0, 1)$.

Conceptually, a firm's pricing problem can be broken into two parts. First, firm $i$'s choice of marginal prices determines contract allocations and hence expected gross social surpluses from serving each type, $S^i_L$ and $S^i_H$. Second, firm $i$'s choice of fixed fees then determines the expected gross utilities offered to each type, $U^i_L$ and $U^i_H$. The differences $\mu^i_\theta \equiv \left(S^i_\theta - U^i_\theta\right)$ are the firm's expected markups on each contract and the profits per customer served.

[29] Exogenous inattention is not essential. All the results already either limit penalty fees directly (Lemma 2 bounds $|p_{3\theta}| < 1/\bar{f}$ ) or impose market conditions that imply optimal penalty fees satisfy the same bound (Propositions 10-11 and Corollary 2). Allowing for endogenous inattention (finite $k$) simply adds an additional bound on penalty fees which must be satisfied: $p_3 \in \left[\underline{p}(c, c), \bar{p}(c, c)\right]$ (see footnote 17). The punch line in Corollary 2 would not be affected however, as this is already stated for $\tau$ sufficiently small.

When choosing contract markups $\mu_L^i$ and $\mu_H^i$ (via the choice of fixed fees) a firm must ensure that it is incentive compatible for consumers to choose the intended contracts. The constraint that type $H$ not choose contract $L$ is the *downward* incentive constraint. The constraint that type $L$ not choose contract $H$ is the *upward* incentive constraint. Optimal pricing varies according to whether the downward or upward constraint is binding and this, in turn, depends on residual demand.

Analogously to the benchmark model, residual demand is given by the function $G_\theta\left(U_\theta^i; U_\theta^{-i}\right)$, which describes firm $i$'s market share in segment $\theta$ given firm $i$'s utility offer $U_\theta^i$ and offers of other firms $U_\theta^{-i}$. Below, I suppress firm superscript $i$ and competing offers $U_\theta^{-i}$ from the notation. When deriving a firm's best response given exogenous residual demand, I assume $G_\theta\left(U_\theta\right)$ is differentiable and $U_\theta + \frac{G_\theta(U_\theta)}{g_\theta(U_\theta)}$ is strictly increasing. This decreasing-marginal-revenue assumption is satisfied in a Hotelling duopoly and ensures that the firm has a uniquely optimal markup in each segment.

An important characteristic of residual demand is the unconstrained optimal markup:

**Definition 4** *The unconstrained optimal markup $\mu_\theta^*$ is the optimal markup for segment $\theta$ given first-best allocations and ignoring ex ante incentive constraints: $\mu_\theta^* = S_\theta^{FB} - \hat{U}_\theta$ where $\hat{U}_\theta \equiv \arg\max_U G_\theta\left(U\right)\left(S_\theta^{FB} - U\right)$.*[30]

Unconstrained optimal markups are those that would be charged under third-degree price discrimination. Residual demand will satisfy one of three cases: (1) $\mu_L^* = \mu_H^*$, (2) $\mu_H^* > \mu_L^*$, or (3) $\mu_H^* < \mu_L^*$. The relative ranking of unconstrained optimal markups determines whether downward or upward incentive constraints bind and, hence, the qualitative nature of optimal pricing.

## 5.2 Pricing with bill-shock regulation (Attentive Case)

Consider a Hotelling duopoly in which transport costs for high and low types satisfy $\tau_H > \tau_L$. This is a natural assumption if high-average-value customers are high-income customers who have a low marginal-value of money and therefore strong brand preferences.[31] Importantly, a consequence of this assumption is that all equilibria are inefficient:

---

[30] Thus, $\mu_\theta^* = \frac{G_\theta(\hat{U}_\theta)}{g_\theta(\hat{U}_\theta)}$ where $\hat{U}_\theta$ uniquely satisfies $S_\theta^{FB} = \hat{U}_\theta + \frac{G_\theta(\hat{U}_\theta)}{g_\theta(\hat{U}_\theta)}$.

[31] It is standard to assume in empirical demand estimation that those who have high marginal value for vertical attributes such as quantity or quality also have high marginal value for horizontal attributes such as brand due to income effects. For example, the seminal automobile demand estimation by Berry, Levinsohn and Pakes (1995) makes this assumption. See the working paper version (Grubb, 2012) for analysis of the alternative case $\tau_H \leq \tau_L$. The important assumption is that $\tau_H \neq \tau_L$, as otherwise there would be no price discrimination.

**Proposition 9** *Hotelling duopoly Equilibrium (with bill-shock regulation): If $\tau_H > \tau_L$ then: (1) All equilibria are inefficient. (2) In all symmetric equilibria, high types receive first-best allocations, while low types' allocations are distorted downwards.*

Inefficient allocations arise for the standard reason in second-degree price-discrimination models: to give one group a discounted markup relative to another, the discount must be accompanied by a distorted allocation to prevent everyone choosing the discounted markup.

The first step to prove Proposition 9 is to characterize a firm's best response given exogenous residual demand (see Lemma 1 in Appendix A). The solution to the firm's problem depends on how unconstrained optimal markups (Definition 4) are ranked across low and high market-segments. When there is no reason to price discriminate ($\mu_L^* = \mu_H^*$) neither ex ante incentive constraint binds and the firm offers a single first-best contract. When market segment $L$ would receive a discounted markup under third-degree price discrimination ($\mu_L^* < \mu_H^*$) the downward incentive constraint is binding, contract $H$ is first best, and marginal prices on contract $L$ are above marginal cost, distorting allocations downwards. When market segment $H$ would receive a discounted markup under third-degree price discrimination ($\mu_L^* > \mu_H^*$) the reverse is true: the upward incentive constraint is binding, contract $L$ is first best, and marginal prices on contract $H$ are below marginal cost, distorting allocations upwards.

The second step to prove Proposition 9 is to show that unconstrained optimal markups have the same relative ranking ($\mu_H^* > \mu_L^*$) in a Hotelling duopoly as transport costs ($\tau_H > \tau_L$). This result is intuitive, recalling that with a single market segment equilibrium markups would equal the transport costs. Given this result, Proposition 9 follows from a firm's best response characterized in Appendix A Lemma 1.

### 5.3 Pricing without bill-shock regulation (Inattentive case)

I now characterize equilibrium pricing when consumers are inattentive and respond only to the expected marginal price because they do not keep track of past usage. I focus on the case relevant to fairly competitive markets, in which unconstrained optimal markups are not too different. It is striking that, in contrast to the attentive case, firms can charge somewhat different markups to different market segments without distorting allocations. This leads to the result that bill-shock regulation will reduce welfare in fairly-competitive markets.

Let $v^s_{t\theta\hat{\theta}}$ be the optimal time $t$ consumption threshold of a sophisticated inattentive consumer of type $\theta$ who chooses contract $\hat{\theta}$, and let $v^s_{t\theta} = v^s_{t\theta\theta}$. The first-order conditions for $v^s_{t\theta\hat{\theta}}$ are a natural

extension of equations (5)-(6):

$$v^s_{1\theta\hat{\theta}} = p_{1\hat{\theta}} + p_{3\hat{\theta}}\left(1 - F_\theta(v^s_{2\theta\hat{\theta}})\right), \tag{20}$$

$$v^s_{2\theta\hat{\theta}} = p_{2\hat{\theta}} + p_{3\hat{\theta}}\left(1 - F_\theta(v^s_{1\theta\hat{\theta}})\right). \tag{21}$$

A sophisticated inattentive consumer $\theta$ who chooses contract $\hat{\theta}$ earns expected gross utility

$$U_{\theta\hat{\theta}} = -p_{0\hat{\theta}} + \sum_{t\in\{1,2\}} \int_{v^s_{t\theta\hat{\theta}}}^1 (v - p_{t\theta})\, dF_\theta(v) - p_{3\hat{\theta}}\left(1 - F_\theta(v^s_{1\theta\hat{\theta}})\right)\left(1 - F_\theta(v^s_{2\theta\hat{\theta}})\right), \tag{22}$$

and for $\hat{\theta} = \theta$ earns $U_\theta = U_{\theta\theta}$ and generates expected gross surplus

$$S_\theta = \sum_{t\in\{1,2\}} \int_{v^s_{t\theta}}^1 (v - c)\, dF_\theta(v). \tag{23}$$

It is useful to reframe the firm's problem in two ways. First, think of the firm choosing offered utility levels $U_\theta$ rather than fixed fees $p_{0\theta}$, which are then determined by equation (22) evaluated at $\hat{\theta} = \theta$. Second, think of the firm first choosing consumer thresholds $v^s_{t\theta}$, so that $p_{t\theta}$ are determined by equations (20)-(21), and then choosing the best penalty fee $p_{3\theta}$ which makes $v^s_{t\theta}$ globally (rather than just locally) incentive compatible. Then the firm's problem can be written as:

$$\max_{\substack{U_L, v^s_{1L}, v^s_{2L}, p_{3L} \\ U_H, v^s_{1H}, v^s_{2H}, p_{3H}}} (1-\beta) G_L(U_L)\left(S_L(v^s_{1L}, v^s_{2L}) - U_L\right) + \beta G_H(U_H)\left(S_H(v^s_{1H}, v^s_{2H}) - U_H\right)$$

s.t. $U_H \geq U_{HL}$ (downward IC), $U_L \geq U_{LH}$ (upward IC),

and $v^s_{1L}$, $v^s_{2L}$, $v^s_{1H}$, and $v^s_{2H}$ are incentive compatible,[32]

where $U_{HL}$, $U_{LH}$, $S_L$, and $S_H$ are given by equations (22) and (23) and $p_{1\theta}$, $p_{2\theta}$, and $p_{0\theta}$ are given by equations (20)-(22) evaluated at $\hat{\theta} = \theta$ for $\theta \in \{L, H\}$.

Notice that only offered utilities $U_\theta$ and consumer thresholds $v^s_{t\theta}$ enter the objective function directly. Penalty fees $p_{3\theta}$ only affect profits via the incentive constraints. If pricing is symmetric ($v^s_{1\theta} = v^s_{2\theta}$) increasing $p_{3\theta}$ above zero initially weakly relaxes both upward and downward ex ante incentive constraints, and hence weakly increases profits (Lemma 2, Appendix A). Raising $p_{3\theta}$ is eventually infeasible however, as for $p_{3\theta}$ sufficiently large, the chosen calling thresholds will not be implementable (Lemma 3, Appendix A).

Building on the preceding insights, Proposition 10 characterizes a firm's best response, treating residual demand $G_\theta(U_\theta)$ as exogenous. The crucial difference relative to the attentive case is that both upward and downward incentive constraints may be slack even when unconstrained optimal markups differ, as long as the difference is not too large. Proposition 10 characterizes the best response in this case, as it is the one that will be relevant to fairly competitive markets.

[32] $\{v^s_{1\theta}, v^s_{2\theta}\} \in \arg\max_{x_1,x_2} \left\{\sum_{t\in\{1,2\}} \int_{x_t}^1 (v - p_{t\theta})\, dF_\theta(v) - p_{3\hat{\theta}}(1 - F_\theta(x_1))(1 - F_\theta(x_2))\right\}$ for $\theta \in \{L, H\}$.

**Proposition 10** *Firm Best Response (without bill-shock regulation): There exist strictly positive constants $X_L, X_H > 0$ such that the following holds. If exogenous residual demand curves $\{G_L(U_L), G_H(U_H)\}$ satisfy $\mu_H^* - \mu_L^* \in [-X_L, X_H]$ then: (1) Optimal contracts implement first-best allocations ($v_{1L}^s = v_{2L}^s = v_{1H}^s = v_{2H}^s = c$) and unconstrained optimal mark-ups ($\mu_L^*$, $\mu_H^*$). (2) If $\mu_H^* \neq \mu_L^*$ then surprise penalty fees are charged but not disclosed at the point-of-sale.*

Comparing Proposition 10 with the best-response results described in Section 5.2 shows the underlying insight of the third main result: the combination of surprise penalty fees and consumer inattention can be both profitable and socially valuable by reducing allocative distortions due to price discrimination when unconstrained optimal markups differ across consumer segments but are not too different. In the attentive problem, contracts implement first-best allocations only for the knife-edge case $\mu_L^* = \mu_H^*$. With inattentive consumers this is no longer true. Slack ex ante incentive constraints and first-best allocations are a feature for $(\mu_H^* - \mu_L^*)$ in an interval around zero because penalty fees relax incentive constraints when consumers are inattentive.[33]

To elaborate, penalty fees relax incentive constraints whether or not consumers are attentive. However, penalty fees always distort second period allocations of attentive consumers because $p_2$ and $p_2 + p_3$ cannot both be equal to marginal cost if $p_3 > 0$. The important difference when consumers are inattentive is that allocations can be efficient despite positive surprise penalty fees because inattentive consumers respond to the expected marginal price $v_{2\theta}$ rather than the actual marginal price $p_2$ or $p_2 + p_3$.

For more intuition, suppose that $\mu_H^* > \mu_L^*$. If consumers are attentive, firms cannot induce first-best allocations and charge low types a discounted markup. First-best allocations require marginal-cost pricing for every unit on every contract. With identical marginal prices on all contracts, all consumers would choose the lowest fixed fee and pay the same markup. To discount low types' markup, firms must combine a discounted fixed fee with higher marginal prices that distort quantity choices. The striking result for inattentive consumers is that this is no longer the case for small discounts. First-best allocations only require that expected marginal prices equal marginal cost and could, for instance, be implemented by offering $p_{1\theta} = p_{2\theta} = 0$ and $p_{3\theta} = c/(1 - F_\theta(c))$ if $c/(1 - F_\theta(c)) < 1/\bar{f}_\theta$. As high types pay penalty fees more often than low types, these contracts involve lower penalty fees on the high contract to achieve the same expected marginal price. Moreover, high types are willing to pay a higher increase in the fixed-fee for a reduction in the penalty fee than

[33] Proposition 10 shows that firms can price discriminate between inattentive consumers without paying information rents. Prior work provides examples in which firms can price discriminate between consumers with biased beliefs without paying information rents. For instance, see the illustrative example in the introduction of Eliaz and Spiegler (2008) and footnote 25 in Grubb (2009).

are low types. As a result, the low-type contract can offer a discounted markup without distorting allocations or attracting high types.

To illustrate the preceding intuition, consider the following example:

**Example 2** *For low types, $v_t \sim U[0,10]$, and for high types, $v_t \sim U[0,15]$. Marginal cost is $c = 5$.*
*Contract L: Free first unit and a \$10 penalty: $p_{1L} = p_{2L} = 0$ and $p_{3L} = 10$.*
*Contract H: Free first unit and a \$7.5 penalty: $p_{1H} = p_{2H} = 0$ and $p_{3H} = 7.5$.*

Both contracts in Example 2 are efficient for their intended consumers. For low types who choose contract $L$, the optimal consumption threshold is equal to marginal cost: $v_L^s = 5$. At this consumption threshold, low types will purchase with probability $1/2$ in each period so that, conditional on purchasing in the current period, the expected marginal price is $1/2$ times the penalty fee or $(1/2)\,10 = 5$. Similarly, for high types who choose contract $H$, the optimal consumption threshold is also equal to marginal cost: $v_H^s = 5$. At this consumption threshold, high types will purchase with probability $2/3$ in each period. Thus the lower penalty fee is exactly offset by a higher purchase probability so that expected marginal price is the same: $(2/3)\,7.5 = 5$. Moreover, as contract $H$ has a lower penalty fee than contract $L$, the firm can charge a higher fixed fee on contract $H$. In fact, the fixed fee on contract $H$ can be up to \$1 higher while maintaining incentive compatibility. The \$1 difference in fixed fees corresponds to a higher markup on contract $H$: $\mu_H = \mu_L + 1/6$. Thus inattention and surprise penalty fees allow the firm to charge different markups without distorting allocations.

The preceding paragraphs focus on the case in which unconstrained optimal markups satisfy $(\mu_H^* - \mu_L^*) \in [-X_L, X_H]$, for which Proposition 10 shows that allocations are first best. This is the relevant case given sufficient competition because competition compresses unconstrained optimal markups. Firms with stiff competitors never want to charge high types too large a premium over low types because they would lose business to other firms. Thus, if consumers are inattentive, fairly competitive markets will satisfy $(\mu_H^* - \mu_L^*) \in [-X_L, X_H]$ and yield efficient outcomes with surprise penalty fees. To state the result formally, let $\tau_H = \tau H$, $\tau_L = \tau L$, $H > L > 0$, and $\tau > 0$, so that $\tau$ parameterizes the degree of competition. Also define $\bar{f} \equiv \max_{v,\theta} f_\theta(v)$.

**Proposition 11** *Hotelling duopoly Equilibrium (without bill-shock regulation): If $\tau > 0$ is sufficiently small then: Without bill-shock regulation, in the unique (up to penalty fees) symmetric pure-strategy equilibrium, all customers are served, allocations are first best, and mark-ups are $\mu_\theta = \tau_\theta$. Moreover, surprise penalty fees are charged but not disclosed at the point-of-sale. If $c/(1 - F_L(c)) < 1/\bar{f}$ then the set of equilibrium prices includes $p_{1\theta}^i = p_{2\theta}^i = 0$ and $p_{3\theta}^i = c/(1 - F_\theta(c))$.*

The intuition for Proposition 11 is as follows. In equilibrium, unconstrained optimal markups are closely related to transportation costs. Thus, in fairly-competitive markets when $\tau$ is small, and hence the difference between $\tau_H$ and $\tau_L$ is also small, unconstrained optimal markups will satisfy $(\mu_H^* - \mu_L^*) \in [-X_L, X_H]$. (Constants $X_L$ and $X_H$ are independent of $\tau$.) As a result, Proposition 10 implies that, absent bill-shock regulation, firms will price discriminate without distorting allocations. Equilibrium without bill-shock regulation is efficient because competition ensures firms only want to charge high types a slightly higher markup and this can be achieved using only surprise penalty fees.[34]

### 5.4 Consequences of bill-shock regulation

By comparing Propositions 9 and 11, Corollary 2 completes the third main result. The combination of surprise penalty fees and consumer inattention are socially valuable and bill-shock regulation is counterproductive whenever markets are fairly competitive.

**Corollary 2** *Consider a Hotelling duopoly with transportation costs $\tau_H = \tau H > \tau_L = \tau L > 0$. If $\tau > 0$ is sufficiently small then: (1) Bill-shock regulation would strictly decrease welfare and firm profits. (2) Low types would lose while high types would win.*

Corollary 2 follows by comparing Propositions 9 and 11. Absent bill-shock regulation, Proposition 11 shows equilibrium allocations are efficient for $\tau > 0$ sufficiently small. In contrast, given bill-shock regulation, Proposition 9 implies equilibrium allocations are inefficient for all $\tau > 0$. Thus bill-shock regulation strictly lowers social welfare in fairly competitive markets. It does so because bill-shock regulation eliminates surprise penalty fees from firms' price-discrimination toolbox and forces firms to introduce quantity distortions on contract $L$.

To understand the distributional results, note that, by necessitating quantity distortions, bill-shock regulation makes price discrimination less profitable for firms. Thus firms also respond to bill-shock regulation by charging more similar markups: increasing the markup on contract $L$ and reducing the markup on contract $H$. Low types are hurt both by the quantity distortion and the higher markup on contract $L$ while high types benefit from the markup reduction on contract $H$. Firm market shares are unaffected in equilibrium, but profits are reduced because the loss from reducing markups on contract $H$ exceed the gains from raising markups on contract $L$ by a factor

[34]The efficiency result in Proposition 11 relies on several assumptions. For instance, if consumers were risk averse, penalty fees would have an inherent social cost. If some consumers were attentive, then those consumers would make inefficient consumption choices. Nevertheless, relaxing these assumptions slightly would only lead to a small deviation from efficiency because the model is continuous.

of $H/L$. This is because $L$ types are more price-sensitive, so on the margin it is more expensive to raise markups on contract $L$ in terms of market share.[35]

A caveat to Corollary 2 is that the magnitude of the effects may be small. For instance, consider equal proportions of the consumer types described in Example 2, the same marginal cost, and a Hotelling duopoly with $\tau_L = 1/6$ and $\tau_H = 1/3$ (Example 3 in the online appendix). As should be expected, the welfare consequences of bill-shock regulation are less than the difference in transportation costs $\Delta\tau = 1/6$, which underlies all price-discrimination related distortions. Bill-shock regulation lowers utility to low types by about 22% of $\Delta\tau$, creates a deadweight-loss of about 5% of $\Delta\tau$ for each low type, and raises utility to high types by about 16% of $\Delta\tau$. Changes in industry profits are smaller. In the limit as a market becomes perfectly competitive and $\Delta\tau$ goes to zero, the magnitude of the welfare changes described by Corollary 2 approach zero. Thus in sufficiently competitive markets the harmful effects of bill-shock regulation should not be a concern. Moreover, the impact of regulation becomes ambiguous when there is sufficient market power. Both surprise penalty fees and quantity distortions are useful tools for price discrimination. In some cases (including fairly-competitive markets) they are substitutes and regulation that eliminates surprise increases quantity distortions. In other cases they are complementary and the reverse is true.[36]

# 6 Conclusion

For electricity tariffs, health insurance, or loyalty discounts, contract design should be cognizant of consumer inattention. If inattentive consumers are sophisticated then they will optimally respond to expected marginal prices. If inattentive consumers are naive then they underestimate the probability of paying surprise penalty fees (via endogenous overconfidence) and overestimate the probability of redeeming loyalty discounts or perks (via endogenous underconfidence).

There is little scope for bill-shock regulation to affect consumers when they are homogeneous. When consumers are heterogeneous, however, bill-shock regulation may be beneficial or harmful, depending on the nature of heterogeneity. When some consumers are attentive but more are naively inattentive, bill-shock regulation increases social welfare, benefits naive consumers by ending cross subsidization, and can benefit all consumers by stiffening competition. In contrast, if sophisticated consumers with heterogeneous forecasts of their future demand are inattentive, then bill-shock

[35]Shifts in markups in each segment are already inversely weighted by shares of each segment $\beta$ and $(1-\beta)$ since the shares reflect the cost of distorting that segment. Thus the difference in price sensitivity drives the difference in relative profit changes, rather than relative segment sizes.

[36]See Corollary 3 in working paper version (Grubb, 2012).

regulation will be socially harmful in fairly-competitive markets, although possibly by only a small amount.

It is interesting to consider how other types of heterogeneity might affect the results. A richer model could include sophisticated as well as attentive and naive consumers in the same market. In this case sophisticated consumers would always receive a separate contract with marginal cost pricing. In a Hotelling duopoly with equal transportation costs across consumer types, the sophisticated contract would be independent of that for attentive and naive consumers. If $\tau_a < \tau_s$, however, penalty fees or loyalty discounts on the attentive contract could be useful not only for exploiting naive consumers but also for price discrimination. They could help induce sophisticates to pay a premium for marginal-cost pricing to avoid penalty fees they know they can't keep track of.

The results suggest that regulators should require bill-shock alerts for services such as overdraft protection that are not differentially priced to sort consumers across contracts. This is particularly true for the case of overdraft fees in which (1) there is evidence of heterogeneity in attention (Stango and Zinman, 2014), (2) the limited impact of the Federal Reserve Board's opt-in regulation on overdraft-fee revenues suggests it is a poor substitute for bill-shock regulation, and (3) high opt-in rates are consistent with naivete. However, regulators should be more cautious about cases including the FCC's recent bill-shock agreement in which penalty fees do help sort consumers across contracts.

# A Additional Details Referenced in Section 5

## A.1 Price discrimination with bill-shock regulation (Attentive Case)

I characterize a firm's best response given the model of Section 5.1, attentive consumers, and exogenous residual demand $G_\theta(U_\theta)$ that is differentiable and satisfies $U_\theta + \frac{G_\theta(U_\theta)}{g_\theta(U_\theta)}$ strictly increasing.[37] I assume consumers are attentive because this will be the outcome if bill-shock regulation is imposed.

Let $v^a_{\theta\hat{\theta}}$ be the optimal first-period consumption-threshold of an attentive consumer of type $\theta$ who chooses contract $\hat{\theta}$ and let $v^a_\theta = v^a_{\theta\theta}$. The expression for $v^a_{\theta\hat{\theta}}$ is an extension of equation (4):

$$v^a_{\theta\hat{\theta}} = p_{1\hat{\theta}} + \int_{p_{2\hat{\theta}}}^{p_{2\hat{\theta}}+p_{3\hat{\theta}}} (1 - F_\theta(v))\, dv. \tag{24}$$

An attentive consumer $\theta$ who chooses contract $\hat{\theta}$ earns expected gross utility

$$\begin{aligned} U_{\theta\hat{\theta}} \;=\; & -p_{0\hat{\theta}} + \int_{v^a_{\theta\hat{\theta}}}^{1} \left(v - p_{1\hat{\theta}}\right) dF_\theta(v) \\ & + F_\theta(v^a_{\theta\hat{\theta}}) \int_{p_{2\hat{\theta}}}^{1} \left(v - p_{2\hat{\theta}}\right) dF_\theta(v) + \left(1 - F_\theta(v^a_{\theta\hat{\theta}})\right) \int_{p_{2\hat{\theta}}+p_{3\hat{\theta}}}^{1} \left(v - p_{2\hat{\theta}} - p_{3\hat{\theta}}\right) dF_\theta(v), \end{aligned} \tag{25}$$

and for $\hat{\theta} = \theta$ earns $U_\theta = U_{\theta\theta}$ and generates expected gross social surplus

$$S_\theta = \int_{v^a_\theta}^{1} (v - c)\, dF_\theta(v) + \int_{p_{2\theta}+p_{3\theta}}^{1} (v - c)\, dF_\theta(v) + F_\theta(v^a_\theta) \int_{p_{2\theta}}^{p_{2\theta}+p_{3\theta}} (v - c)\, dF_\theta(v). \tag{26}$$

It is useful to reframe the firm's problem in two ways. First, think of the firm choosing offered utility levels $U_\theta$ so that fixed fees $p_{0\theta}$ are determined by equation (25) evaluated at $\hat{\theta} = \theta$ as function of $U_\theta$. Second, think of the firm choosing a consumer's first-period threshold $v^a_\theta$ rather than marginal price $p_{1\theta}$. Given a choice of $v^a_\theta$, it is necessary for $p_{1\theta}$ to satisfy equation (24) evaluated at $\hat{\theta} = \theta$. The firm's problem can then be written as:

$$\max_{\substack{U_L, v^a_L, p_{2L}, p_{3L} \\ U_H, v^a_H, p_{2H}, p_{3H}}} (1-\beta)\, G_L(U_L)\left(S_L(v^a_L, p_{2L}, p_{3L}) - U_L\right) + \beta G_H(U_H)\left(S_H(v^a_H, p_{2H}, p_{3H}) - U_H\right)$$

$$\text{s.t. } U_H \geq U_{HL} \text{ (downward IC) and } U_L \geq U_{LH} \text{ (upward IC),}$$

[37] An alternative assumption, that $G_\theta(U_\theta) = 1$ if $U_\theta \geq 0$ and $G_\theta(U_\theta) = 0$ otherwise, would capture a monopolist serving consumers with zero outside option. Lemma 1 and Proposition 10 hold under this alternative (see working paper version (Grubb, 2012)). I assume there are $T = 2$ sub-periods when quantity choices are made after a contract is signed. Given attentive consumers, $T = 1$, and $G_\theta(U_\theta) = 1_{U_\theta \geq 0}$, my model would coincide with Courty and Li (2000), which models airline-ticket refund-contracts. Given attentive consumers, $T \geq 1$, and $G_\theta(U_\theta) = 1_{U_\theta \geq 0}$, my model would nearly be a special case of the problem studied by Pavan, Segal and Toikka (2014). However, because I assume period-zero types are discrete rather than continuous, Pavan et al.'s (2014) results do not apply, and conditional independence of values does not lead to a repetition of the Courty and Li (2000) solution. I allow for heterogeneous outside-options so that I can move beyond monopoly pricing and analyze imperfect competition.

where $U_{HL}$, $U_{LH}$, $S_L$, and $S_H$ are given by equations (25) and (26) and $p_{1\theta}$ and $p_{0\theta}$ are given by equations (24) and (25) evaluated at $\hat{\theta} = \theta$ for $\theta \in \{L, H\}$.

As outlined in Section 5.1, demand curves fall into one of three categories. Lemma 1 characterizes optimal monopoly contracts for each case and shows that penalty fees are strictly positive on distortionary contracts, irrespective of the direction of the distortion.

**Lemma 1** *Firm Best Response (with bill-shock regulation): Given exogenous residual demand curves $G_L(U_L)$ and $G_H(U_H)$ satisfy decreasing marginal revenue, optimal contracts satisfy:*

1. *If $\mu_L^* = \mu_H^*$, then a single marginal-cost contract with markup $\mu_L^*$ gives first-best allocations.*
2. *If $\mu_H^* > \mu_L^*$, then $H$'s allocation is first best via marginal-cost pricing but $L$'s allocation is distorted downwards: $v_L^a, p_{2L}, p_{2L} + p_{3L} > c$. Penalty fee $p_{3L}$ is strictly positive.*
3. *If $\mu_H^* < \mu_L^*$, then $L$'s allocation is first best via marginal-cost pricing but $H$'s allocation is distorted upwards: $v_H^a, p_{2H}, p_{2H} + p_{3H} < c$. Penalty fee $p_{3H}$ is strictly positive.*

Given case (2) of Lemma 1 ($\mu_H^* > \mu_L^*$), the triple $\{v_L^a, p_{2L}, p_{3L}\}$ satisfies equations (27)-(29), the first-order conditions for marginal prices on contract $L$:

$$v_L^a = c + \int_{p_{2L}}^{p_{2L}+p_{3L}} (v-c) f_L(v)\, dv + \frac{-\partial\Pi}{\partial U_H} \frac{F_L(v_L^a) - F_H(v_{HL}^a)}{(1-\beta) G_L(U_L) f_L\left(v_L^a\right)}, \tag{27}$$

$$p_{2L} = c + \frac{F_H(v_{HL}^a)}{F_L\left(v_L^a\right)} \frac{-\partial\Pi}{\partial U_H} \frac{F_L(p_{2L}) - F_H(p_{2L})}{(1-\beta) G_L(U_L) f_L(p_{2L})}, \tag{28}$$

$$p_{2L} + p_{3L} = c + \frac{1 - F_H(v_{HL}^a)}{1 - F_L\left(v_L^a\right)} \frac{-\partial\Pi}{\partial U_H} \frac{F_L(p_{2L}+p_{3L}) - F_H(p_{2L}+p_{3L})}{(1-\beta) G_L(U_L) f_L(p_{2L}+p_{3L})}, \tag{29}$$

where $v_{HL}^a = v_L^a + \int_{p_{2L}}^{p_{2L}+p_{3L}} (F_L(v) - F_H(v))\, dv$.

Given case (3) of Lemma 1 ($\mu_H^* < \mu_L^*$), the triple $\{v_H^a, p_{2H}, p_{3H}\}$ satisfies equations (30)-(32), the first-order conditions for marginal prices on contract $H$:

$$v_H^a = c + \int_{p_{2H}}^{p_{2H}+p_{3H}} (v-c) f_H(v)\, dv - \frac{-\partial\Pi}{\partial U_L} \frac{F_L(v_{LH}^a) - F_H(v_H^a)}{\beta G_H(U_H) f_H\left(v_H^a\right)}, \tag{30}$$

$$p_{2H} = c - \frac{F_L(v_{LH}^a)}{F_H\left(v_H^a\right)} \frac{-\partial\Pi}{\partial U_L} \frac{F_L(p_{2H}) - F_H(p_{2H})}{\beta G_H(U_H) f_H(p_{2H})}, \tag{31}$$

$$p_{2H} + p_{3H} = c - \frac{1 - F_L(v_{LH}^a)}{1 - F_H\left(v_H^a\right)} \frac{-\partial\Pi}{\partial U_L} \frac{F_L(p_{2H}+p_{3H}) - F_H(p_{2H}+p_{3H})}{\beta G_H(U_H) f_H(p_{2H}+p_{3H})}, \tag{32}$$

where $v_{LH}^a = v_L^a - \int_{p_{2L}}^{p_{2L}+p_{3L}} (F_L(v) - F_H(v))\, dv$.

For intuition, consider each case in turn. (1) If $\mu_H^* = \mu_L^*$ then firms have no desire to price discriminate, which means that a single marginal cost contract is optimal. (2) If $\mu_H^* > \mu_L^*$ then

marginal prices are distorted above marginal cost on contract $L$ to relax the downward incentive constraint and dissuade $H$ types from choosing discounted contract $L$. If both types choose the same contract, the penalty fee is always more likely to be paid by the high type. Thus the penalty fee on contract $L$ is positive to increase the positive price distortion, targeting it towards deviating high types but away from low types who actually choose contract $L$. (3) If $\mu_H^* < \mu_L^*$ then marginal prices are distorted below marginal cost on contract $H$ to relax the upward incentive constraint and dissuade $L$ types from choosing discounted contract $H$. In this case, the penalty fee on contract $H$ is positive to decrease the negative price distortion, targeting it towards deviating low types but away from high types who actually choose the contract. (More intuition is in the online appendix.)

Lemma 1 can explain the use of penalty fees but not the use of *surprise* penalty fees. Importantly, Lemma 1 shows that allocations are first best only when unconstrained optimal markups are identical for both types. This knife-edge efficiency-result is analogous to findings by Armstrong and Vickers (2001) and Rochet and Stole (2002) in a static rather than sequential screening context.

### A.2 Price discrimination without bill-shock regulation (Inattentive Case)

The following results apply to the model of Section 5.3. For more intuition see the online appendix.

**Lemma 2** *Let $v_{1\theta}^s = v_{2\theta}^s = v_\theta^s$ and $|p_{3\theta}| < 1/\bar{f}$ for $\theta \in \{L, H\}$. Then: (1) Increasing $p_{3L}$ weakly relaxes the downward incentive constraint without affecting the upward incentive constraint. (2) Increasing $p_{3H}$ weakly relaxes the upward incentive constraint without affecting the downward incentive constraint.*

**Lemma 3** *For each interior consumption threshold pair $\{v_{1\theta}^s, v_{2\theta}^s\} \in (0,1)^2$, there exist finite constants $\underline{p}_{3\theta}, \bar{p}_{3\theta} > 0$ such that $\{v_{1\theta}^s, v_{2\theta}^s\}$ is not implementable if $p_{3\theta} \notin [-\underline{p}_{3\theta}, \bar{p}_{3\theta}]$.*

To see why Lemma 3 is true, consider a fixed threshold pair $(v_{1\theta}^s, v_{2\theta}^s)$ the firm would like to implement. Base marginal charges $p_{1\theta}$ and $p_{2\theta}$ must be chosen to satisfy the necessary conditions for local incentive compatibility in equations (5)-(6). In that case, raising $p_{3\theta} > 0$ sufficiently high will mean that the consumer prefers the threshold pair $(0, 1)$, as this strategy earns a first-purchase subsidy $p_{1\theta} < 0$ for sure but never pays the penalty fee. Similarly, lowering $p_{3\theta} < 0$ to be sufficiently negative will mean that the consumer prefers the threshold pair $(0, 0)$ which ensures the loyalty discount is always collected.

## B Proofs

Note that in the proofs I use the notation $p_4 = p_2 + p_3$.

## B.1 Proof of Proposition 1

The second period strategy is optimal by inspection. Given $v_2^a = p_2 + q_1 p_3$, the expected gross utility from choosing first-period threshold $v_1^a$ and its derivative are:

$$U_a\left(v_1^a\right) = -p_0 + \int_{v_1^a}^1 \left(v_1 - p_1 + \int_{p_4}^1 \left(v_2 - p_4\right) f\left(v_2\right) dv_2\right) f\left(v_1\right) dv_1 + F\left(v_1^a\right) \int_{p_2}^1 \left(v_2 - p_2\right) f\left(v_2\right) dv_2,$$

$$\frac{dU_a}{dv_1^a} = f\left(v_1^a\right) \left(-v_1^a + p_1 + \int_{p_2}^{p_4} \left(v_2 - p_2\right) f\left(v_2\right) dv_2 + \left(1 - F\left(p_4\right)\right) p_3\right).$$

The first-order condition, $dU_a/dv_1^a = 0$, yields $v_1^a = p_1 + \left(1 - F\left(p_4\right)\right) p_3 + \int_{p_2}^{p_4} \left(v - p_2\right) f\left(v\right) dv$. Integrating this expression by parts yields equation (4). Moreover, this identifies the global maximum since for $v_1^a > p_1 + \int_{p_2}^{p_4} \left(v_2 - p_2\right) f\left(v_2\right) dv_2 + \left(1 - F\left(p_4\right)\right) p_3$, $dU_a/dv_1^a < 0$ and vice-versa.

## B.2 Proof of Proposition 2

**(1) Equations (5)-(6) are necessary:** A feasible inattentive strategy is a pair of functions $\{b_1\left(v_1\right), b_2\left(v_2\right)\}$ which describe a purchase probability for each valuation $v_t$ to be implemented at all $t > 0$ independently of past usage. The consumer chooses these functions to maximize $U\left(b_1, b_2\right)$:

$$U\left(b_1, b_2\right) = -p_0 + \sum_{t \in \{1,2\}} \int_0^1 \left(v - p_t\right) b_t\left(v\right) dF\left(v\right) - p_3 \left(\int_0^1 b_1\left(v\right) dF\left(v\right)\right) \left(\int_0^1 b_2\left(v\right) dF\left(v\right)\right).$$

Pointwise derivatives with respect to $b_t\left(v\right)$ are: $dU\left(b_1, b_2\right)/db_1\left(v\right) = f\left(v\right) \left(v - p_1 - p_3 \int_0^1 b_2\left(v\right) dF\left(v\right)\right)$, and $dU\left(b_1, b_2\right)/db_2\left(v\right) = f\left(v\right) \left(v - p_2 - p_3 \int_0^1 b_1\left(v\right) dF\left(v\right)\right)$. These derivatives show by inspection that a threshold hold rule is optimal in each period and that equations (5)-(6) are necessary conditions for these thresholds to be optimal (up to the fact that all thresholds above one are equivalent and all thresholds below zero are equivalent).

**(2) If $|p_3| < 1/\bar{f}$ then equations (5)-(6) are sufficient:** Note that we can substitute equation (6) into equation (5) to obtain $v_1^s = p_1 + \left(1 - F\left(p_2 + \left(1 - F\left(v_1^s\right)\right) p_3\right)\right) p_3$. The derivative of the right hand side of this expression with respect to $v_1^s$ is $p_3^2 f\left(v_1^s\right) f\left(p_2 + \left(1 - F\left(v_1^s\right)\right) p_3\right)$, or $p_3^2 f\left(v_1^s\right) f\left(v_2^s\right)$. Thus a sufficient condition for a unique fixed point is $p_3^2 f\left(v_1^s\right) f\left(v_2^s\right) < 1$, for which $|p_3| < 1/\bar{f}$ is in turn sufficient, where $\bar{f} = \max_v f\left(v\right)$. Therefore $|p_3| < 1/\bar{f}$ implies that equations (5)-(6) are sufficient as well as necessary for the pair $\{v_1^s, v_2^s\}$ to be optimal. (This conclusion relies on knowing that an optimal pair $\{v_1^s, v_2^s\}$ exists. This follows because consumer utility is a continuous function of $\{v_1^s, v_2^s\}$ and varies only on the compact set $[0,1]^2$.)

**(3) If $|p_3| < 1/\bar{f}$ and $p_1 = p_2$ then $v_1^s = v_2^s$:** If prices are symmetric ($p_1 = p_2 = p$) then equations (5)-(6) have a symmetric solution $v_1^s = v_2^s = v^s$ satisfying $v^s = p + \left(1 - F\left(v^s\right)\right) p_3$. (A fixed point always exists within $\left[\min\left\{p, p + p_3\right\}, \max\left\{p, p + p_3\right\}\right]$ as the right hand side of the

preceding expression lies within this interval.) Therefore, when there is a unique solution (as for $|p_3| < 1/\bar{f}$ ) it must be symmetric.

### B.3 Proof of Proposition 3

Firm $i$'s profits can be written as $\Pi^i = G\left(U^i; U^{-i}\right)(S^i - U^i)$. For any fixed utility offer $U^i$, profits are maximized by choosing marginal prices $p_1^i$, $p_2^i$, and $p_3^i$ to achieve first-best gross surplus, while adjusting the fixed fee $p_0^i$ to keep $U^i$ constant. This is true independent of regulation. The offered gross utility $U^i$ is set via the fixed fee $p_0^i$ to balance rent extraction versus participation, as in a basic monopoly pricing problem. Firms' gross utility offers are independent of regulation, which implies that matching between firms and consumers and transportation costs are also independent of regulation. Given attentive consumers and continuously-distributed taste-shocks, $p_1^i = p_2^i = c$ and $p_3^i = 0$ are the unique marginal prices which achieve $S^{FB}$. Given inattentive consumers and continuously-distributed taste-shocks, any marginal prices which implement $v_1^s = v_2^s = c$ are optimal. By Proposition 3, these include all marginal prices which satisfy $\left|p_3^i\right| < 1/\bar{f}$ and equations (5)-(6) for $v_1^s = v_2^s = c$. These correspond to the three-part tariffs described in the proposition.

### B.4 Proof of Proposition 4

Period 1 behavior follows from naivete and Proposition 1. Conditional on choosing to be inattentive, the expected payoff at time 2 as a function of the period 2 consumption choice $q_2$ is

$$U_n(q_2) = -p_0 + \int_{v_1^a}^{1} (v_1 - p_1) f(v_1)\, dv_1 + q_2 \left(v_2 - p_2 - p_3 \left(1 - F\left(v_1^a\right)\right)\right).$$

By inspection, second period purchase is optimal when $v_2 \geq p_2 + p_3\left(1 - F\left(v_1^a\right)\right)$. Choosing to be attentive yields expected utility $U_a - k$. Thus naive consumers are inattentive if $U_a - U_n = \Delta_n \leq k$.

### B.5 Proof of Proposition 5

Inattention follows from $\Delta_n \leq k$ and Proposition 4. Contract overvaluation ($\Delta_n > 0$) for $p_3 \neq 0$ follows from inspection of equation (11) after recognizing from equation (10) that $p_3 \neq 0$ implies $v_2^n \neq p_2 + p_3$. Let $Q = q_1 + q_2$. The true probability distribution over $Q$ is $\Pr(Q = 0) = F\left(v_1^a\right) F\left(v_2^n\right)$, $\Pr(Q = 1) = F\left(v_1^a\right)\left(1 - F\left(v_2^n\right)\right) + \left(1 - F\left(v_1^a\right)\right) F\left(v_2^n\right)$, and $\Pr(Q = 2) = \left(1 - F\left(v_1^a\right)\right)\left(1 - F\left(v_2^n\right)\right)$. However, consumers believe the distribution to be $\Pr^*(Q = 0) = F\left(v_1^a\right) F(p_2)$, $\Pr^*(Q = 1) = F\left(v_1^a\right)(1 - F(p_2)) + \left(1 - F\left(v_1^a\right)\right) F(p_4)$, and $\Pr^*(Q = 2) = \left(1 - F\left(v_1^a\right)\right)(1 - F(p_4))$. For $p_3 > 0$, equation (10) implies $p_2 < v_2^n < p_4$ and thus $\Pr^*(Q = 2) < \Pr(Q = 2)$ and $\Pr^*(Q = 0) < \Pr(Q = 0)$. Thus consumers underestimate the likelihood of paying penalties and are overconfident ($H^*(Q)$ crosses $H(Q)$ from below at $Q = 1$). For $p_3 < 0$, equation (10) implies $p_2 > v_2^n > p_4$ and

thus $\Pr^*(Q=2) > \Pr(Q=2)$ and $\Pr^*(Q=0) > \Pr(Q=0)$. Thus consumers overestimate the likelihood of receiving a loyalty discount and are underconfident ($H^*(Q)$ crosses $H(Q)$ once from above at $Q=1$).

### B.6 Proof of Proposition 6

Proposition 6 follows from the discussion in the text except for three points. First, I outline the derivation of equation (16): Subtracting equation (9) from equation (7) and simplifying yields (11). From equation (13), $dp_2/dp_3 = -(1-F(v_1^a))$. Calculating $d\Delta_n/dp_3 = \partial\Delta_n/\partial p_3 - (1-F(v_1^a))\,\partial\Delta_n/\partial p_2$ and simplifying yields equation (16).

Second, the text implicitly assumes that the optimal $v_1^a \in (0,1)$. This is consistent with the result $v_1^a = c$ as $c \in (0,1)$ by assumption. Moreover, it is readily apparent that the predicted contract dominates any other with $v_1^a \notin (0,1)$. For $v_1^a \notin (0,1)$, $\Delta_n = 0$. Therefore, any other such contract could yield profits no higher than $G_n(U_a)\left(S^{FB} - U_a\right)$, which is strictly less than $G_n(U_a)\left(S^{FB} - U_a + k\right)$. Therefore $v_1^a \in (0,1)$ is optimal.

Third, the firm finds it strictly optimal not to disclose whether $p_3$ is applicable at the point of sale. This follows because disclosure substitutes for attention and makes naivete irrelevant. Thus, given disclosure, the optimal contract is marginal cost pricing with $p_3 = 0$ (Proposition 3). This yields profits $G_n(U_a)\,k$ lower for any utility offer $U_a$ so is not optimal.

### B.7 Proof of Proposition 7

Solving equation (7) for the fixed fee $p_0$ yields:

$$p_0 = -U_a + \int_{v_1^a}^{1}\left(v - p_1 + \int_{p_2+p_3}^{1}(v - p_2 - p_3)\,f(v)\,dv\right)f(v)\,dv + F(v_1^a)\int_{p_2}^{1}(v-p_2)\,f(v)\,dv. \quad (33)$$

By Proposition 6, optimal consumption thresholds are $v_1^a = v_2^n = c$. Substituting equations (12)-(13) and values $v_1^a = v_2^n = c$ into equation (33) yields $p_0$ as a function of $U_a$ and $p_3$: $p_0(U_a, p_3)$. Evaluating at $p_3 = 0$ yields $p_0(U_a, 0) = S^{FB} - U_a$. For a given $U_a$, Proposition 6 implies that $p_0\left(U_a, p^{\min}(c,c)\right)$ is the optimal loyalty-discount-contract fixed-fee and $p_0(U_a, p^{\max}(c,c))$ is the optimal penalty-fee-contract fixed-fee.

Differentiating $p_0(U_a, p_3)$ yields $dp_0(U_a,p_3)/dp_3 = (1-F(c))\left((1-F(p_4)) + F(c)\left(F(p_4) - F(p_2)\right)\right)$, or equivalently $dp_0(U_a,p_3)/dp_3 = (1-F(c))\left((1-F(p_2)) + (1-F(c))\left(F(p_2) - F(p_4)\right)\right)$, where $p_2 = c - (1-F(c))\,p_3$ and $p_4 = c + F(c)\,p_3$. Inspection of these expressions shows that the assumption $c \in (0,1)$ implies $dp_0(U_a,p_3)/dp_3 > 0$. Therefore $p_0\left(U_a, p^{\min}(c,c)\right) < p_0(U_a,0) = S^{FB} - U_a < p_0(U_a, p^{\max}(c,c))$ follows from $p^{\min}(c,c) < 0 < p^{\max}(c,c)$. Moreover, the differences

$p_0(U_a, 0) - p_0\left(U_a, p^{\min}(c,c)\right)$ and $p_0(U_a, p^{\max}(c,c)) - p_0(U_a, 0)$ are independent of $U_a$. Thus for $U_a$ sufficiently close to $S^{FB}$, $p_0\left(U_a, p^{\min}(c,c)\right) < 0 < p_0(U_a, p^{\max}(c,c))$.

### B.8 Proof of Corollary 1

Part (1) and the fact that allocations are efficient with or without regulation both follow immediately from comparing Propositions 3 and 6 given the observation that bill-shock regulation makes both sophisticated and naive inattentive consumers indistinguishable from attentive consumers. Given strict full-market-coverage, Hotelling duopoly yields symmetric markups equal to $\tau$. As I assume strict full-market-coverage with or without regulation,[38] this means that market shares (and hence transportation costs) and profits are the same with or without regulation. Combined with the fact that allocations are efficient with or without regulation, this result implies part (2).

### B.9 Proof of Proposition 8

The proof proceeds in 4 steps. First, I compute first and second order derivatives of the profit function ignoring the constraint $\Delta_n \leq k$. Then I turn to each of the proposition's three results.

**Derivation of first-order conditions:** Equation (18) specifies the firm's objective, where $\Delta_n$, $S_n$, and $S_a$ are functions of $\{p_3, v_1^a, v_2^n, p_2\}$ given by equations (11), (14), and (19) respectively. Equation (13) determines $p_2$ and its derivatives $dp_2/dx$ with respect to $x \in \{p_3, v_1^a, v_2^n\}$. Derivatives of firm profits with respect to $x \in \{p_3, v_1^a, v_2^n\}$ are computed as $d\Pi/dx = \partial\Pi/\partial x + (\partial\Pi/p_2)(dp_2/dx)$:

$$\frac{d\Pi}{dp_3} = (1 - F(v_1^a)) F(v_1^a) \left( \begin{array}{c} -\alpha_a G_a(U_a)((c-p_2) f(p_2) + (p_4 - c) f(p_4)) \\ +\alpha_n G_n(U_a)(F(p_4) - F(p_2)) \end{array} \right), \tag{34}$$

$$\frac{d\Pi}{dv_1^a} = f(v_1^a) \left( \begin{array}{c} \alpha_a G_a(U_a)\left(-v_1^a + c + \int_{p_2}^{p_4} (v-c) f(v)\, dv - p_3 X\right) \\ +\alpha_n G_n(U_a)\left(-v_1^a + c + \int_{p_2}^{p_4} (v - p_2 - p_3 F(v_1^a)) f(v)\, dv\right) \end{array} \right), \tag{35}$$

$$\begin{aligned} \frac{d\Pi}{dv_2^n} = & -\alpha_a G_a(U_a) F(v_1^a)(p_2 - c) f(p_2) + (1 - F(v_1^a))(p_4 - c) f(p_4) \\ & + \alpha_n G_n(U_a) f(v_2^n) \left( -v_2^n + c + \frac{1}{f(v_2^n)} \left( \begin{array}{c} (F(p_4) - F(v_2^n)) \\ -F(v_1^a)(F(p_4) - F(p_2)) \end{array} \right) \right). \end{aligned} \tag{36}$$

[38] Hotelling duopoly involves strict full-market-coverage if and only if, at a markup of $\tau$, gross expected utility $U_a = S_n - \tau + \Delta_n$ exceeds transport cost $\tau/2$ for the median consumer. Absent regulation, this requires $\tau < \frac{2}{3}\left(S^{FB} + k\right)$. With regulation, this requires $\tau < \frac{2}{3} S^{FB}$. Thus my assumption of strict full-market-coverage with or without regulation corresponds to the assumption $0 \leq \tau < \frac{2}{3} S^{FB}$.

Similarly, I compute $d^2\Pi/dp_3^2$, expressed compactly using notation $W$ and $Y$ below:

$$\begin{aligned}\frac{d^2\Pi}{dp_3^2} &= \left(1-F\left(v_1^a\right)\right)F\left(v_1^a\right)\left(\alpha_n G_n\left(U_a\right)Y-\alpha_a G_a\left(U_a\right)\left(Y+W\right)\right), \qquad (37)\\ W &= F\left(v_1^a\right)\left(p_4-c\right)f'\left(p_4\right)+\left(1-F\left(v_1^a\right)\right)\left(p_2-c\right)f'\left(p_2\right),\\ Y &= F\left(v_1^a\right)f\left(p_4\right)+\left(1-F\left(v_1^a\right)\right)f\left(p_2\right).\end{aligned}$$

**Result (1), $\alpha_n > \alpha_a$ implies $p_3^i \neq 0$ for at least one firm $i \in \{A, B\}$:** Suppose not and that $\alpha_n > \alpha_a$ but $p_3^i = 0$ at both firms $i \in \{A, B\}$. $\Delta_n = 0$ for $p_3 = 0$ so the constraint $\Delta_n \leq k$ is slack in a neighborhood of $p_3 = 0$. Thus first-order conditions $d\Pi/dv_1^a = d\Pi/dv_2^n = 0$ must hold. At $p_3 = 0$, $v_2^n = p_2 = p_4$ and the derivatives of firm profits with respect to $v_1^a$ and $v_2^n$ from equations (35)-(36) reduce to $d\Pi/dv_1^a = (c - v_1^a) f(v_1^a)(\alpha_a G_a(U_a) + \alpha_n G_n(U_a))$ and $d\Pi/dv_2^n = (c - v_2^n) f(v_2^n)(\alpha_a G_a(U_a) + \alpha_n G_n(U_a))$. Hence the firms must also both choose $v_1^a = v_2^n = c$. In that case, the first and second derivatives of firm profits with respect to $p_3$, from equations (34) and (37), reduce to $d\Pi/dp_3 = 0$ and $d^2\Pi/dp_3^2 = (1 - F(c)) F(c) f(c)(\alpha_n G_n(U_a) - \alpha_a G_a(U_a))$. Given the assumed full market coverage in both naive and attentive segments, $\alpha_n > \alpha_a$ implies that for at least one firm it holds that $\alpha_n G_n(U_a) > \alpha_a G_a(U_a)$. For this firm, $d\Pi/dp_3 = 0$ and $d^2\Pi/dp_3^2 > 0$. This implies that $p_3 = 0$ is a local minimum and hence the firm is not best-responding. This contradiction proves the result.

**Result (2), Bill-shock regulation increases social welfare:** This follows immediately from part (1) and the benchmark result. Without regulation part (1) implies that at least one firm offers $p_3^i \neq 0$. At this firm, attentive consumers will make inefficient choices in period 2 meaning that social surplus is strictly below first best. As bill-shock regulation is assumed a perfect substitute for attention, the benchmark result (Proposition 3) implies it will increase social surplus to first-best.

**Result (3), Distributional results for equal transport cost case:** The equal transport cost assumption ($\tau = \tau_a = \tau_n$) implies that each firm $i$ attracts the same fraction of attentive consumers as naive consumers. Thus, the firm's objective function in equation (18) may be re-written factoring out $G^i\left(U_a^i\right) = G_a^i\left(U_a^i\right) = G_n^i\left(U_a^i\right)$:

$$\max_{\substack{U_a^i, v_1^{a,i}, v_2^{n,i}, p_3^i \\ \text{such that } \Delta_n \leq k}} G_a\left(U_a^i\right)\left(-U_a^i + \alpha_a S_a\left(v_1^{a,i}, v_2^{n,i}, p_3^i\right) + \alpha_n\left(S_n\left(v_1^{a,i}, v_2^{n,i}\right) + \Delta_n\left(v_1^{a,i}, v_2^{n,i}, p_3^i\right)\right)\right).$$

Inspection of the objective function shows that marginal prices $(v_1^{a,i}, v_2^{n,i}, p_3^i)$ are chosen to maximize the sum $\alpha_a S_a\left(v_1^{a,i}, v_2^{n,i}, p_3^i\right) + \alpha_n\left(S_n\left(v_1^{a,i}, v_2^{n,i}\right) + \Delta_n\left(v_1^{a,i}, v_2^{n,i}, p_3^i\right)\right)$ subject to $\Delta_n \leq k$ but independent of the utility offer $U_a^i$ or of the other firm's choices. As a result, in any equilibrium, both firms choose the same marginal prices $(v_1^a, v_2^n, p_3)$, implementing the same

$$S^* \equiv \max_{\substack{v_1^a, v_2^n, p_3 \\ \text{such that } \Delta_n \leq k}} \alpha_a S_a\left(v_1^a, v_2^n, p_3\right) + \alpha_n\left(S_n\left(v_1^a, v_2^n\right) + \Delta_n\left(v_1^a, v_2^n, p_3\right)\right). \qquad (38)$$

Firm's therefore choose $U_a^i$ to maximize $\Pi^A = G_a\left(U_a^i\right)\left(S^* - U_a^i\right)$. As $\tau$ is assumed sufficiently small for strict full-market-coverage, it is then a standard result that, in the unique equilibrium, firms split the market with symmetric offers $U_a^A = U_a^B = S^* - \tau$ and profits $\Pi^A = \Pi^B = \tau/2$.

Under bill-shock regulation, firms each earn $\tau/2$ and consumers receive $S^{FB} - 5\tau/4$ (where $S^{FB}$ is gross first-best surplus, $\tau$ is the firm markup, and the extra $\tau/4$ is the average transportation cost). Optimality of marginal prices implies $S^* \geq S^{FB}$. Moreover, given $\alpha_n > \alpha_a$ the inequality is strict, $S^* > S^{FB}$, following the proof of result (1).[39] Therefore attentive agents lose the difference $S^* - S^{FB}$ from bill-shock regulation while firm profits are unaffected. Because social welfare strictly increases (part (2)), naive consumers must be better off.

### B.10 Solution to Example 1

Equilibria in the example are computed numerically. Firm $i$ best responses to firm $j$'s utility offer $U_a^j$ are calculated by searching numerically over the choice variables $(U_a^i, v_1^{a,i}, v_2^{n,i}, p_3^i)$ to maximize profits. Then best response functions are iterated until they converge on an equilibrium. The search is simplified substantially by recognizing several features of the example: (1) First, the optimal choice of $U_a^i$ is a straightforward analytical function of $(v_1^{a,i}, v_2^{n,i}, p_3^i)$ and $U_a^j$. (2) Second, the profit function is symmetric: $\Pi\left(U_a, v_1^a, v_2^n, p_3\right) = \Pi\left(U_a, \hat{v}_1^a, v_2^n, \hat{p}_3\right)$ for $\hat{p}_3 = -p_3$ and $\hat{v}_1^a = 1 - v_1^a$. Thus it is sufficient to consider only $p_3 \geq 0$, recognizing that any solutions found will have a twin satisfying $\hat{p}_3 = -p_3$ and $\hat{v}_1^a = 1 - v_1^a$. (3) Third, within the range $p_2, p_4 \in (0,1)$, it is locally optimal for $v_1^a = v_2^n = 1/2$, $p_3 = 0$ if $\alpha_a G_a\left(U_a\right) \geq \alpha_n G_n\left(U_a\right)$, and $\Delta_n\left(p_3\right) = k$ if $\alpha_a G_a\left(U_a\right) < \alpha_n G_n\left(U_a\right)$. (While other local maxima for $p_2 < 0$ or $p_4 > 1$ must be checked, they turn out not to be relevant in this example.) Full details are available from the author upon request.

### B.11 Proof of Lemma 1

Note, in the proof I write the firm's problem as a choice of marginal prices $p_{2\theta}$ and $p_{4\theta}$ rather than $p_{2\theta}$ and $p_{3\theta}$, where $p_{4\theta} = p_{2\theta} + p_{3\theta}$.

**Case I**: The result for $\mu_H^* = \mu_L^*$ follows because the optimal solution when both IC constraints are relaxed is a single marginal-cost contract with markup $\mu_L^*$.

**Case II**. Assume $\mu_H^* > \mu_L^*$. Relax the upward incentive constraint $U_L \geq U_{LH}$ (IC-L).

**(1) Claim**: IC-L slack implies marginal cost pricing ($v_H^a = p_{2H} = c$ and $p_{3H} = 0$) and first-best allocation for the high type. **Proof**: Suppose not. Then setting $\{v_H^a, p_{2H}, p_{4H}\}$ equal to $\{c, c, c\}$

[39] The proof of part (1) shows that, holding constant $p_1 = p_2 = c$ and any fixed $U_a$, profits achieve a local minimum in $p_3$ at $p_3 = 0$. For equal transportation costs, the maximand of equation (38) must also achieve a local minimum in $p_3$ at the same point, where it equals $S^{FB}$. Therefore its maximum, $S^*$, must be strictly higher.

while keeping $U_H$ constant keeps IC-H and participation unaffected without violating IC-L since it has been relaxed. However, it increases surplus and hence profit from type $H$ - a contradiction.

**(2) Claim**: The downward incentive constraint $U_H \geq U_{HL}$ (IC-H) binds with equality: $U_H = U_{HL}$. **Proof**: Suppose that IC-H were slack ($U_H > U_{HL}$). Then there would be marginal cost pricing $P_\theta(\mathbf{q}) = p_{0\theta} + c(q_1 + q_2)$, IC-H reduces to $p_{0H} \leq p_{0L}$, markups would equal fixed fees $\mu_\theta = p_{0\theta}$, and markups would be at the unconstrained optima $\mu_\theta = \mu_\theta^*$ (see below). Thus IC-H implies $\mu_H^* \leq \mu_L^*$, which is a contradiction, so IC-H must bind. Moreover, IC-H will bind with equality because the decreasing marginal revenue assumption, $U_\theta + \frac{G_\theta(U_\theta)}{g_\theta(U_\theta)}$ increasing, implies profits are quasi-concave in $U_\theta$.

Note that the claim above that $\mu_\theta = \mu_\theta^*$ follows in three steps. (1) First, $U_L$ and $U_H$ must be at an unconstrained local profit maximum because IC-H is slack and IC-L is relaxed. (2) Second, the fact that profits are quasi-concave in $U_H$ implies that the unconstrained local maximum must be the unconstrained global maximum. (3) Third, $\mu_\theta = \mu_\theta^*$ follows from the definition of $\mu_\theta^*$ given marginal cost pricing.

**(3)** The downward incentive constraint $U_H \geq U_{HL}$ (IC-H) is convenient to re-express as $U_H \geq U_L + (U_{HL} - U_L)$. Let $Z = (U_{HL} - U_L)$. Integrating by parts, equation (25) reduces to:

$$U_{\theta\hat{\theta}} = -p_{0\hat{\theta}} + \int_{v^a_{\theta\hat{\theta}}}^1 \left(v - p_{1\hat{\theta}}\right) dF_\theta(v) + F_\theta\left(v^a_{\theta\hat{\theta}}\right) \int_{p_{2\hat{\theta}}}^1 (1 - F_\theta(v))\, dv + \left(1 - F_\theta\left(v^a_{\theta\hat{\theta}}\right)\right) \int_{p_{4\hat{\theta}}}^1 (1 - F_\theta(v))\, dv. \tag{39}$$

Thus the expression for $Z$ can be re-written as:

$$\begin{aligned} Z \;=\; & \int_{v^a_{HL}}^1 (v - p_{1L})\, f_H(v)\, dv + \int_{p_{2L}}^1 \left(F_H(v^a_{HL})(1 - F_H(v)) - F_L(v^a_L)(1 - F_L(v))\right) dv \\ & - \int_{v^a_L}^1 (v - p_{1L})\, f_L(v)\, dv + \int_{p_{4L}}^1 \left((1 - F_H(v^a_{HL}))(1 - F_H(v)) - (1 - F_L(v^a_L))(1 - F_L(v))\right) dv, \end{aligned} \tag{40}$$

where from equation (24) evaluated at $\hat{\theta} = \theta = L$,

$$p_{1L} = v^a_L - \int_{p_{2L}}^{p_{4L}} (1 - F_L(v))\, dv, \tag{41}$$

and substituting this into equation (24) evaluated at $\{\theta, \hat{\theta}\} = \{H, L\}$,

$$v^a_{HL} = v^a_L + \int_{p_{2L}}^{p_{4L}} (F_L(v) - F_H(v))\, dv. \tag{42}$$

Given (1) and (2), the firm's problem can be reduced to:

$$\max_{U_L, v^a_L, p_{2L}, p_{4L}} \left\{ \begin{array}{c} (1 - \beta)\, G_L(U_L)\left(S_L(v^a_L, p_{2L}, p_{4L}) - U_L\right) \\ + \beta G_H\left(U_H(U_L, v^a_L, p_{2L}, p_{4L})\right)\left(S^{FB}_H - U_H(U_L, v^a_L, p_{2L}, p_{4L})\right) \end{array} \right\}$$

where $S_L$ is given by equation (26), $U_H\left(U_L, v_L^a, p_{2L}, p_{4L}\right) = U_L + Z\left(v_L^a, p_{2L}, p_{4L}\right)$, and $Z$ is characterized by equations (40)-(42).

For the remainder of the proof, I suppress subscript "$L$" from $p_{1L}$, $p_{2L}$, and $p_{4L}$. The derivative of firm profits with respect to any $x \in \{v_L^a, p_2, p_4\}$ is $d\Pi/dx = (\partial\Pi/\partial S_L)(dS_L/dx)+(\partial\Pi/\partial U_H)(dU_H/dx)$. The derivatives of $S_L$ are derived from equation (26). Next, $\partial U_H/\partial y = \partial Z/\partial y$ for each $y \in \{v_L^a, p_2, p_4, p_1, v_{HL}^a\}$. Moreover, the definition $Z = U_{HL} - U_L$ and the envelope condition imply that $\partial Z/\partial v_{HL}^a = \partial Z/\partial v_L^a = 0$. Thus for any $x \in \{v_L^a, p_2, p_4\}$, $dU_H/dx = \partial Z/\partial x + (\partial Z/\partial p_1)(dp_1/dx)$, where $\partial Z/\partial x$ and $\partial Z/\partial p_1$ are derived from equation (40) and $dp_1/dx$ is derived from equation (41). Putting these pieces together gives the following derivatives of firm profits:

$$\frac{d\Pi}{dv_L^a} = (1-\beta)\, G_L\left(U_L\right) f_L\left(v_L^a\right)\left(\int_{p_2}^{p_4} (v-c)\, f_L\left(v\right) dv - \left(v_L^a - c\right)\right) + \frac{\partial \Pi}{\partial U_H}\left(F_H\left(v_{HL}^a\right) - F_L\left(v_L^a\right)\right),$$

$$\frac{d\Pi}{dp_2} = -\left(1-\beta\right) G_L\left(U_L\right) F_L\left(v_L^a\right)\left(p_2 - c\right) f_L\left(p_2\right) + \frac{\partial \Pi}{\partial U_H} F_H\left(v_{HL}^a\right)\left(F_H\left(p_2\right) - F_L\left(p_2\right)\right), \tag{43}$$

$$\frac{d\Pi}{dp_4} = -\left(1-\beta\right) G_L\left(U_L\right)\left(1 - F_L\left(v_L^a\right)\right)\left(p_4 - c\right) f_L\left(p_4\right) + \frac{\partial \Pi}{\partial U_H}\left(1 - F_H\left(v_{HL}^a\right)\right)\left(F_H\left(p_4\right) - F_L\left(p_4\right)\right), \tag{44}$$

which can be set equal to zero and rearranged to derive equations (27)-(29) in Appendix A.1.

**(4) Claim**: (a) $p_2, p_4 > c$, (b) $p_4 > p_2$, and (c) $v_{HL}^a \geq v_L^a > c$.

**(a) Claim**: $p_2, p_4 > c$: **Proof**: The fact that IC-H is binding implies that $\partial\Pi/\partial U_H < 0$. By inspection of equations (28)-(29), first-order stochastic dominance (FOSD) implies $p_2 \geq c$ and $p_4 \geq c$. Moreover, $p_2 \neq c$ and $p_4 \neq c$ because FOSD is strict at $c$. Therefore $p_2, p_4 > c$.

**(b) Claim**: $p_4 > p_2$. **Proof**: Suppose not and $p_2 \geq p_4$. If $p_2 = p_4$ then $v_{HL}^a = v_L^a$ by equation (42). Given $p_2 = p_4$ and $v_{HL}^a = v_L^a$ equations (28)-(29) imply $F_H\left(v_L^a\right) = F_L\left(v_L^a\right)$. Hence equation (27) implies $v_L^a = c$, which contradicts $F_H\left(v_L^a\right) = F_L\left(v_L^a\right)$ given strict FOSD at $c$. Thus $p_2 \neq p_4$.

If $p_2 > p_4$ then $v_{HL}^a < v_L^a$ by equation (42) and FOSD. (Note that the inequality is strict because equations (28)-(29) and $p_2, p_4 > c$ imply that $F_H\left(p_2\right) < F_L\left(p_2\right)$ and $F_H\left(p_4\right) < F_L\left(p_4\right)$. Therefore, by continuity of $F_H$ and $F_L$, $\int_{p_4}^{p_2}\left(F_L\left(v\right) - F_H\left(v\right)\right) dv > 0$.) Given $v_{HL}^a < v_L^a$, $F_H$ strictly increasing and FOSD imply $F_H\left(v_{HL}^a\right) < F_H\left(v_L^a\right) \leq F_L\left(v_L^a\right)$. Therefore it holds that $\frac{1-F_H\left(v_{HL}^a\right)}{1-F_L\left(v_L^a\right)} > 1 > \frac{F_H\left(v_{HL}^a\right)}{F_L\left(v_L^a\right)}$, which in turn implies that $\int_a^b \frac{d\Pi}{dp_4} > 0$ follows from $\int_a^b \frac{d\Pi}{dp_2} \geq 0$ for any $b > a$ given equations (43)-(44). Now the fact that $p_2 > p_4$ is optimal implies that $\int_{p4}^{p2} \frac{d\Pi}{dp_2} \geq 0$ so it must also be true that $\int_{p4}^{p2} \frac{d\Pi}{dp_4} > 0$, contradicting optimality of $p_4$. Therefore $p_4 > p_2$.

**(c) Claim**: $v_{HL}^a \geq v_L^a > c$. **Proof**: First, $v_{HL}^a \geq v_L^a$ follows from equation (42), $p_4 > p_2$, and FOSD. Second, given $F_L\left(v_L^a\right) - F_H\left(v_{HL}^a\right) \geq 0$ and $p_4 > p_2 > c$, equation (27) implies that $v_L^a > c$. Thus, it is sufficient to show that $F_L\left(v_L^a\right) - F_H\left(v_{HL}^a\right) \geq 0$. Suppose not and $F_L\left(v_L^a\right) - F_H\left(v_{HL}^a\right) < 0$.

Then $\frac{1-F_H\left(v_{HL}^a\right)}{1-F_L\left(v_L^a\right)} < 1 < \frac{F_H\left(v_{HL}^a\right)}{F_L\left(v_L^a\right)}$ and, by a similar comparison of derivatives in equations (43)-(44) as made above in part (b), it follows that $p_2 > p_4$, which is a contradiction.

**(5) Claim**: IC-L is satisfied. **Proof**: The final step is to show that the relaxed IC-L constraint is satisfied. This follows from the fact that quantities are monotonic in the ex ante signal: $q_{t,H}\left(v^t\right) \geq q_{t,L}\left(v^t\right)$. To show that IC-L is satisfied, it is sufficient to show that $U_H - U_{HL} + U_L - U_{LH} \geq 0$. Because $U_H = U_{HL}$, this implies $U_L \geq U_{LH}$ and therefore that IC-L is satisfied.

Solving equation (24) for $p_{1\hat{\theta}}$ yields $p_{1\hat{\theta}} = v_{\theta\hat{\theta}} - \int_{p_{2\hat{\theta}}}^{p_{4\hat{\theta}}} \left(1 - F_\theta\left(v\right)\right) dv$. Substituting this expression for $p_{1\hat{\theta}}$ into equation (39) and applying the fact that $\int_a^1 \left(x - a\right) dF\left(x\right) = \int_a^1 \left(1 - F\left(x\right)\right) dx$ yields a simplified expression for $U_{\theta\hat{\theta}}$:

$$U_{\theta\hat{\theta}} = -p_{0\hat{\theta}} + \int_{v_{\theta\hat{\theta}}}^1 \left(1 - F_\theta\left(v\right)\right) dv + \int_{p_{2\hat{\theta}}}^1 \left(1 - F_\theta\left(v\right)\right) dv. \tag{45}$$

It follows from equation (45) that the difference $U_\theta - U_{\theta\hat{\theta}}$ is

$$U_{\theta\theta} - U_{\theta\hat{\theta}} = p_{0\hat{\theta}} - p_{0\theta} + \int_{v_\theta}^{v_{\theta\hat{\theta}}} \left(1 - F_\theta\left(v\right)\right) dv + \int_{p_{2\theta}}^{p_{2\hat{\theta}}} \left(1 - F_\theta\left(v\right)\right) dv. \tag{46}$$

After substituting $p_{2H} = v_H^a = v_{LH}^a = c$ (following Claim 1), the sum of interest is then

$$U_H - U_{HL} + U_L - U_{LH} = \int_c^{p_{2L}} \left(F_L\left(v\right) - F_H\left(v\right)\right) dv + \int_c^{v_L^a} \left(F_L\left(v\right) - F_H\left(v\right)\right) dv + \int_{v_L^a}^{v_{HL}^a} \left(1 - F_H\left(v\right)\right) dv. \tag{47}$$

Noting from Claim 4 that $p_{2L} > c$ and $v_{HL}^a \geq v_L^a > c$, FOSD implies that all three terms are positive. Therefore IC-L is satisfied.

**Case III**. The result for $\mu_H^* < \mu_L^*$ follows by a symmetric argument, where I start by relaxing IC-H and showing that IC-L must bind with equality. There is one additional step in this proof. In particular, in Case III the final step remains to show that $U_H - U_{HL} + U_L - U_{LH} > 0$ but the symmetric analog of equation (47) is

$$U_H - U_{HL} + U_L - U_{LH} = \int_{p_{2H}}^c \left(F_L\left(v\right) - F_H\left(v\right)\right) dv + \int_{v_H^a}^c \left(F_L\left(v\right) - F_H\left(v\right)\right) dv - \int_{v_{LH}^a}^{v_H^a} \left(1 - F_L\left(v\right)\right) dv.$$

The first two terms are positive but (unlike in Case II) the third is negative. Thus an additional step is needed to show that $U_H - U_{HL} + U_L - U_{LH} > 0$. In particular, the third term can be usefully bounded by $-\int_{v_{LH}^a}^{v_H^a} \left(1 - F_L\left(v\right)\right) dv \geq -\left(v_H^a - v_{LH}^a\right)\left(1 - F_L\left(v_{LH}^a\right)\right)$. Substituting the symmetric analog of equation (42), $v_{LH}^a - v_H^a = -\int_{p_{2H}}^{p_{4H}} \left(F_L\left(v\right) - F_H\left(v\right)\right) dv$, this is equivalent to $-\int_{v_{LH}^a}^{v_H^a} \left(1 - F_L\left(v\right)\right) dv \geq -\left(1 - F_L\left(v_{LH}^a\right)\right) \int_{p_{2H}}^{p_{4H}} \left(F_L\left(v\right) - F_H\left(v\right)\right) dv$. Thus

$$\begin{aligned} U_H - U_{HL} + U_L - U_{LH} \quad \geq \quad & \int_{p_{4H}}^c \left(F_L\left(v\right) - F_H\left(v\right)\right) dv + \int_{v_H^a}^c \left(F_L\left(v\right) - F_H\left(v\right)\right) dv \\ & + F_L\left(v_{LH}^a\right) \int_{p_{2H}}^{p_{4H}} \left(F_L\left(v\right) - F_H\left(v\right)\right) dv. \end{aligned}$$

Noting that the symmetric analog to Claim 4 is that $p_{2H} < p_{4H} < c$ and $v^a_{LH} \leq v^a_H < c$, FOSD implies that all three terms on the right-hand side are positive. Therefore IC-H is satisfied.

### B.12 Proof of Proposition 9

**Preliminary result**: A useful result not included in Lemma 1 is that $\mu^{*A}_H > \mu^{*A}_L$ implies $\mu^A_H > \mu^A_L$: Suppose not and $\mu^{*A}_H > \mu^{*A}_L$ but $\mu^A_H \leq \mu^A_L$. By Lemma 1, $\mu^{*A}_H > \mu^{*A}_L$ implies that $L$'s allocation is distorted downwards below first best and hence $S^A_L < S^{FB}_L$. Therefore, a strictly profitable deviation would be to offer a single marginal-cost contract with markup $\mu = (\max\{\mu^A_H, \mu^{*A}_L\} + \min\{\mu^A_L, \mu^{*A}_H\})/2$. Ignoring the IC constraints, raising $S^A_L$ to $S^{FB}_L$ holding markups fixed strictly raises market share in segment $L$. Then changing markups to $\mu$ moves $\mu^A_\theta$ weakly closer to $\mu^*_\theta$ for $\theta \in \{L, H\}$, thereby weakly raising profits in each segment as allocations are first best and profits are quasi-concave in $U_\theta$. Finally, the deviation contract is trivially incentive compatible. Therefore $\mu^A_H > \mu^A_L$. By symmetric argument, $\mu^{*A}_H < \mu^{*A}_L$ implies $\mu^A_H < \mu^A_L$.

**Proof of proposition:** Assuming $\tau_L$ and $\tau_H$ are sufficiently small for strict full-market-coverage in equilibrium ensures that every customer strictly prefers to purchase from one of the two firms. In this case, firm A's best response utility offer $U^A_\theta$ is always within an open interval for which residual demand from consumers of type $\theta$ is $G_\theta\left(U^A_\theta\right) = \frac{1}{2\tau_\theta}\left(U^A_\theta - U^B_\theta + \tau_\theta\right)$. Note that firm A's profits are linear in firm B's offer $U^B_\theta$ and hence firm A's expected profits and best response depend only on firm B's expected offer $E\left[U^B_\theta\right] = \bar{U}^B_\theta$. Given $G_\theta\left(U^A_\theta\right) = \frac{1}{2\tau_\theta}\left(U^A_\theta - \bar{U}^B_\theta + \tau_\theta\right)$, the definition of $\mu^*_\theta$ implies that $\mu^{*A}_\theta = (\bar{\mu}^B_\theta + \tau_\theta)/2 + (S^{FB}_\theta - \bar{S}^B_\theta)/2$.

**(1)** All equilibria are inefficient: Suppose not, and in equilibrium allocations are efficient. Then $p^B_{3\theta} = 0$ and $p^B_{1\theta} = p^B_{2\theta} = c$. Incentive compatibility implies $p^B_{0L} = p^B_{0H}$ and therefore $\mu^B_H = \mu^B_L = \mu^B$. These statements hold for any offer in B's mixed strategy. Thus $\mu^{*A}_\theta = \frac{1}{2}\left(\bar{\mu}^B + \tau_\theta\right)$, which implies $\mu^{*A}_L < \mu^{*A}_H$, and by Lemma 1 A's best response includes an inefficient contract.

**(2)** Allocations: In a symmetric equilibrium, either $\mu^*_L = \mu^*_H$, $\mu^*_L < \mu^*_H$, or $\mu^*_L > \mu^*_H$ must hold for both firms. Thus it is sufficient to show that $\tau_H > \tau_L$ implies $\mu^*_H > \mu^*_L$, as the result then follows from Lemma 1. First, part (1) rules out $\mu^*_L = \mu^*_H$. Second, suppose that $\mu^*_L > \mu^*_H$ for both firms. For all offers in A's best response, it holds that $\mu^A_L > \mu^A_H$ and that (following the proof of Lemma 1) IC-L will bind while IC-H is slack so that $-\partial\Pi^A/\partial U^A_L = \partial\Pi^A/\partial U^A_H > 0$. Moreover, these inequalities hold in expectation if A uses a mixed strategy: $\bar{\mu}^A_L > \bar{\mu}^A_H$ and $-E\left[\partial\Pi^A/\partial U^A_L\right] = E\left[\partial\Pi^A/\partial U^A_H\right] > 0$. In a symmetric equilibrium $\bar{U}^A_\theta = \bar{U}^B_\theta$ so $E\left[\partial\Pi^A/\partial U^A_H\right] = \left(\bar{\mu}^A_H - \tau_H\right)/2\tau_H$ and $E\left[\partial\Pi^A/\partial U^A_L\right] = \left(\bar{\mu}^A_L - \tau_L\right)/2\tau_L$ and hence these inequalities imply $\tau_H < \bar{\mu}^A_H < \bar{\mu}^A_L < \tau_L$, which contradicts $\tau_H > \tau_L$. Thus $\mu^*_H > \mu^*_L$ for both firms and Lemma 1 implies the result.

## B.13 Proof of Lemma 2

Solving equation (22) at $\hat{\theta} = \theta$ for $p_{0\theta}$ and solving equations (20)-(21) at $\hat{\theta} = \theta$ for $p_{t\theta}$ yields

$$\begin{aligned} p_{0\theta} &= -U_\theta + \sum_{t\in\{1,2\}} \int_{v^s_{t\theta}}^1 (v - p_{t\theta})\, dF_\theta(v) - p_{3\theta}\left(1 - F_\theta\left(v^s_{1\theta}\right)\right)\left(1 - F_\theta\left(v^s_{2\theta}\right)\right), && (48)\\ p_{1\theta} &= v^s_{1\theta} - p_{3\theta}\left(1 - F_\theta\left(v^s_{2\theta}\right)\right), && (49)\\ p_{2\theta} &= v^s_{2\theta} - p_{3\theta}\left(1 - F_\theta\left(v^s_{1\theta}\right)\right). && (50) \end{aligned}$$

Substituting equations (48)-(50) into equation (22) yields

$$\begin{aligned} U_{\theta\hat{\theta}} &= U_{\hat{\theta}} + \sum_{t\in\{1,2\}} \int_{v^s_{t\theta\hat{\theta}}}^1 \left(v - v^s_{t\hat{\theta}}\right) dF_\theta(v) - \sum_{t\in\{1,2\}} \int_{v^s_{t\hat{\theta}}}^1 \left(v - v^s_{t\hat{\theta}}\right) dF_{\hat{\theta}}(v) \qquad (51)\\ &\quad -p_{3\hat{\theta}}\left(F_\theta(v^s_{1\theta\hat{\theta}}) - F_{\hat{\theta}}(v^s_{1\hat{\theta}})\right)\left(F_\theta(v^s_{2\theta\hat{\theta}}) - F_{\hat{\theta}}(v^s_{2\hat{\theta}})\right). \end{aligned}$$

By the envelope condition:

$$\frac{d}{dp_{3\hat{\theta}}} U_{\theta\hat{\theta}} = \frac{\partial}{\partial p_{3\hat{\theta}}} U_{\theta\hat{\theta}} = -\left(F_{\hat{\theta}}(v^s_{1\hat{\theta}}) - F_\theta(v^s_{1\theta\hat{\theta}})\right)\left(F_{\hat{\theta}}(v^s_{2\hat{\theta}}) - F_\theta(v^s_{2\theta\hat{\theta}})\right). \qquad (52)$$

By assumption, $v^s_{1\theta} = v^s_{2\theta} = v^s_\theta$. This implies from equations (49)-(50) that $p_{1\theta} = p_{2\theta} = p_\theta = v^s_\theta - p_{3\theta}\left(1 - F_\theta\left(v^s_\theta\right)\right)$. Therefore, given $|p_{3\theta}| < 1/\bar{f}$ for $\bar{f} = \max_{\theta,v} f_\theta(v)$, Proposition 2 implies that $v^s_{1\hat{\theta}} = v^s_{2\hat{\theta}} = v^s_{\hat{\theta}}$ and $v^s_{1\theta\hat{\theta}} = v^s_{2\theta\hat{\theta}} = v^s_{\theta\hat{\theta}}$. As a result, equation (52) simplifies to $dU_{\theta\hat{\theta}}/dp_{3\hat{\theta}} = -(F_{\hat{\theta}}(v^s_{\hat{\theta}}) - F_\theta(v^s_{\theta\hat{\theta}}))^2 \leq 0$.

## B.14 Proof of Lemma 3

Suppose that the firm wishes to implement consumption threshold pair $(v^s_{1\theta}, v^s_{2\theta})$. Then $p_{1\theta}$ and $p_{2\theta}$ must satisfy equations (49)-(50). If the consumer chooses alternative threshold pair $(\hat{v}_{1\theta}, \hat{v}_{2\theta})$, substituting equations (49)-(50) into equation (22) for $\theta = \hat{\theta}$ shows that gross expected utility is

$$U_\theta = -p_{0\theta} + \sum_{t\in\{1,2\}} \int_{\hat{v}_{t\theta}}^1 (v - v^s_{t\theta})\, dF_\theta(v) + p_{3\theta} \begin{pmatrix} \left(1 - F_\theta\left(\hat{v}_{1\theta}\right)\right)\left(1 - F_\theta\left(v^s_{2\theta}\right)\right) \\ +\left(1 - F_\theta\left(v^s_{1\theta}\right)\right)\left(1 - F_\theta\left(\hat{v}_{2\theta}\right)\right) \\ -\left(1 - F_\theta\left(\hat{v}_{1\theta}\right)\right)\left(1 - F_\theta\left(\hat{v}_{2\theta}\right)\right) \end{pmatrix}.$$

Therefore, the benefit from deviating to $(\hat{v}_{1\theta}, \hat{v}_{2\theta})$ relative to the intended $(v^s_{1\theta}, v^s_{2\theta})$ is

$$U_\theta\left(\hat{v}_{1\theta}, \hat{v}_{2\theta}\right) - U_\theta\left(v^s_{1\theta}, v^s_{2\theta}\right) = \sum_{t\in\{1,2\}} \int_{\hat{v}_{t\theta}}^{v^s_{t\theta}} (v - v^s_{t\theta})\, dF_\theta(v) - p_{3\theta}\left(F_\theta\left(\hat{v}_{1\theta}\right) - F_\theta\left(v^s_{1\theta}\right)\right)\left(F_\theta\left(\hat{v}_{2\theta}\right) - F_\theta\left(v^s_{2\theta}\right)\right). \qquad (53)$$

Following equation (53), the benefit from deviating to $(\hat{v}_{1\theta}, \hat{v}_{2\theta}) = (0, 0)$ is

$$U_\theta(0,0) - U_\theta\left(v^s_{1\theta}, v^s_{2\theta}\right) = \sum_{t\in\{1,2\}} \int_0^{v^s_{t\theta}} (v - v^s_{t\theta})\, dF_\theta(v) - p_{3\theta} F_\theta\left(v^s_{1\theta}\right) F_\theta\left(v^s_{2\theta}\right).$$

As $\int_0^{v^s_{t\theta}} (v - v^s_{t\theta})\, dF_\theta(v) \geq -1$, a lower bound from deviating is thus $U_\theta(0,0) - U_\theta(v^s_{1\theta}, v^s_{2\theta}) \geq -2 - p_{3\theta} F_\theta(v^s_{1\theta}) F_\theta(v^s_{2\theta})$. For deviating to be suboptimal the lower bound on the benefits must be non-positive, which requires $p_{3\theta} \geq -2 (F_\theta(v^s_{1\theta}) F_\theta(v^s_{2\theta}))^{-1}$.

Following equation (53), the benefit from deviating to $(\hat{v}_{1\theta}, \hat{v}_{2\theta}) = (0,1)$ is

$$U_\theta(0,1) - U_\theta(v^s_{1\theta}, v^s_{2\theta}) = \int_0^{v^s_{1\theta}} (v - v^s_{1\theta})\, dF_\theta(v) - \int_{v_{2\theta}}^1 (v - v^s_{2\theta})\, dF_\theta(v) + p_{3\theta} F_\theta(v^s_{1\theta}) (1 - F_\theta(v^s_{2\theta})).$$

A lower bound for this benefit is $U_\theta(0,1) - U_\theta(v^s_{1\theta}, v^s_{2\theta}) \geq -1 - \mu + p_{3\theta} F_\theta(v^s_{1\theta}) (1 - F_\theta(v^s_{2\theta}))$. For deviating to be suboptimal the lower bound on the benefits must be non-positive, which requires $p_{3\theta} \leq (1+\mu) (F_\theta(v^s_{1\theta}) (1 - F_\theta(v^s_{2\theta})))^{-1}$.

Therefore the Lemma holds for $\underline{p}_{3\theta} = 2 (F_\theta(v^s_{1\theta}) F_\theta(v^s_{2\theta}))^{-1}$ and $\bar{p}_{3\theta} = (1+\mu) (F_\theta(v^s_{1\theta}) (1 - F_\theta(v^s_{2\theta})))^{-1}$. Note that tighter bounds may be derived by considering deviations to $(1,0)$ and $(1,1)$.

### B.15 Proof of Propositions 10-11

**Definitions:** Before proceeding to the proof, I introduce some notation. First, recall that in Section 5 I have defined $\bar{f}$ as $\bar{f} = \max_{v,\theta} f_\theta(v)$. Second, I define functions $\phi_H(p^{\max})$ and $\phi_L(p^{\max})$ as

$$\phi_H(p^{\max}) \equiv 2 \int_c^{v^s_{HL}} (v - c)\, dF_H(v) + (v^s_{HL} - c) (F_L(c) - F_H(v^s_{HL})), \tag{54}$$

$$\phi_L(p^{\max}) \equiv 2 \int_{v^s_{LH}}^c (c - v)\, dF_H(v) - (c - v^s_{LH}) (F_L(v^s_{LH}) - F_H(c)), \tag{55}$$

where the dependence on $p^{\max}$ arises through $v^s_{HL}$ and $v^s_{LH}$. The terms $v^s_{HL}$ and $v^s_{LH}$ are both functions of $p^{\max}$ defined implicitly by $v^s_{HL} = c + p^{\max}(F_L(c) - F_H(v^s_{HL}))$ and $v^s_{LH} = c - p^{\max}(F_L(v^s_{LH}) - F_H(c))$. The next step is to state and prove a lemma.

**Lemma 4** *Assume that exogenous residual demand curves are $\{G_L(U_L), G_H(U_H)\}$. Assume that markups $\mu_H > 0$ and $\mu_L > 0$ have been chosen such that $\mu_H - \mu_L \in [-\phi_L(p^{\max}_L), \phi_H(p^{\max}_H)]$ for some $p^{\max}_\theta \in [0, 1/\bar{f})$. (1) If thresholds $v^s_{t\theta}$ and penalty fees $p_{3\theta}$ are chosen to maximize firm profits holding chosen markups fixed then both types receive first-best allocations, $v^s_{1L} = v^s_{2L} = v^s_{1H} = v^s_{2H} = c$, and the set of optimal penalty fees includes $p_{3\theta} = p^{\max}_\theta$. (2) If $\mu_H \neq \mu_L$ then surprise penalty fees are charged but not disclosed at the point-of-sale.*

**Proof.** Proof is by construction. Given fixed markups, profits increase in the share of each type attracted $G_\theta(U_\theta)$. The share of type $\theta$ attracted increases in the utility offered to type $\theta$, which is given by $U_\theta = S_\theta(v^s_{1\theta}, v^s_{2\theta}) - \mu_\theta$. Thus, momentarily ignoring incentive constraints, profits would be maximized by choosing $v^s_{t\theta} = c$ and achieving first-best surplus $S_\theta(c,c) = S^{FB}_\theta$ for each type. Part (1) of the result then follows if incentive constraints are satisfied for $v^s_{t\theta} = c$ and $p_{3\theta} = p^{\max}_\theta$.

(A) Following Proposition 2, if $|p_{3\theta}| < 1/\bar{f}$ (for $\bar{f} = \max_{\theta,v} f(v)$) then $v^s_{t\theta} = c$ is incentive compatible for $p_{t\theta}$ chosen to satisfy equations (20)-(21):

$$p_{1\theta} = p_{2\theta} = p_\theta = c - p_{3\theta}\left(1 - F_\theta(c)\right). \tag{56}$$

Moreover, given this symmetric pricing, we know that consumption thresholds will be symmetric even for types choosing the unintended contract. Combining equations (56) and (20)-(21) yields unique characterizations of these consumption thresholds:

$$v^s_{1HL} = v^s_{2HL} = v^s_{HL} = c + p_{3L}\left(F_L(c) - F_H(v^s_{HL})\right), \tag{57}$$

$$v^s_{1LH} = v^s_{2LH} = v^s_{LH} = c + p_{3H}\left(F_H(c) - F_L(v^s_{LH})\right). \tag{58}$$

(B) By equation (51), IC-L and IC-H are given by equations (59) and (60), respectively:

$$\begin{aligned} U_L \geq U_{LH} = U_H &+ \sum_{t\in\{1,2\}} \int_{v^s_{tLH}}^1 (v - v^s_{tH})\, dF_L(v) - \sum_{t\in\{1,2\}} \int_{v^s_{tH}}^1 (v - v^s_{tH})\, dF_H(v) \\ &- p_{3H}\left(F_L(v^s_{1LH}) - F_H(v^s_{1H})\right)\left(F_L(v^s_{2LH}) - F_H(v^s_{2H})\right), \end{aligned} \tag{59}$$

$$\begin{aligned} U_H \geq U_{HL} = U_L &+ \sum_{t\in\{1,2\}} \int_{v^s_{tHL}}^1 (v - v^s_{tL})\, dF_H(v) - \sum_{t\in\{1,2\}} \int_{v^s_{tL}}^1 (v - v^s_{tL})\, dF_L(v) \\ &- p_{3L}\left(F_H(v^s_{1HL}) - F_L(v^s_{1L})\right)\left(F_H(v^s_{2HL}) - F_L(v^s_{2L})\right). \end{aligned} \tag{60}$$

Substituting $v^s_{t\theta} = c$, $v^s_{1HL} = v^s_{2HL} = v^s_{HL}$, and $v^s_{1LH} = v^s_{2LH} = v^s_{LH}$ from step (A) above, these constraints simplify to

$$U_L \geq U_{LH} = \left(U_H - S^{FB}_H\right) + 2\int_{v^s_{LH}}^1 (v - c)\, dF_L(v) - p_{3H}\left(F_H(c) - F_L(v^s_{LH})\right)^2, \tag{61}$$

$$U_H \geq U_{HL} = \left(U_L - S^{FB}_L\right) + 2\int_{v^s_{HL}}^1 (v - c)\, dF_H(v) - p_{3L}\left(F_L(c) - F_H(v^s_{HL})\right)^2. \tag{62}$$

Noting that (1) $2\int_{v^s_{HL}}^1 (v-c)\, dF_H(v) = S^{FB}_H - 2\int_c^{v^s_{HL}} (v-c)\, dF_H(v)$, (2) $2\int_{v^s_{LH}}^1 (v-c)\, dF_L(v) = S^{FB}_L - 2\int_c^{v^s_{LH}} (v-c)\, dF_L(v)$, and (3) that, following the assumption $v^s_{t\theta} = c$, $(S^{FB}_H - U_H) - (S^{FB}_L - U_L) = \mu_H - \mu_L$, IC-L and IC-H simplify further to:

$$(\mu_H - \mu_L) \geq -2\int_c^{v^s_{LH}} (v - c)\, dF_L(v) - p_{3H}\left(F_H(c) - F_L(v^s_{LH})\right)^2 \tag{63}$$

$$(\mu_H - \mu_L) \leq 2\int_c^{v^s_{HL}} (v - c)\, dF_H(v) + p_{3L}\left(F_L(c) - F_H(v^s_{HL})\right)^2. \tag{64}$$

Finally, substituting $p_{3H}\left(F_H(c) - F_L(v^s_{LH})\right) = v^s_{LH} - c$ and $p_{3L}\left(F_L(c) - F_H(v^s_{HL})\right) = v^s_{HL} - c$ (from equations (57)-(58)) into the right side of equations (63)-(64) shows IC-L and IC-H are

equivalent to $(\mu_H - \mu_L) \geq -\phi_L(p_L^{\max})$ and $(\mu_H - \mu_L) \leq \phi_H(p_H^{\max})$, respectively, when $p_{3L} = p_L^{\max}$ and $p_{3H} = p_H^{\max}$. This completes the proof of Lemma 4 part (1).

(C) Part (2) of Lemma 4 follows because if $p_{3L} = p_{3H} = 0$, we have $v_{HL}^s = v_{LH}^s = c$ (from equations (57)-(58)) and IC-L and IC-H reduce to $(\mu_H - \mu_L) = 0$. Thus $\mu_H \neq \mu_L$ and $v_{t\theta}^s = c$ cannot be supported without penalty fees. Choosing $v_{t\theta}^s \neq c$ for some $t$ and $\theta$ would strictly reduce $U_\theta = S_\theta - \mu_\theta$ below $S_\theta^{FB} - \mu_\theta$ and lead to lower share $G_\theta(U_\theta)$ and lower profits $G_\theta(U_\theta)\mu_\theta$ from that segment. Thus using choosing $p_{3L} = p_{3H} = 0$ is not optimal for $\mu_H \neq \mu_L$. Moreover, disclosing penalty fees is not optimal either, as doing so necessarily induces inefficient consumption, which is costly for the same reasons. ■

**Proof of Proposition 10:** Define $X_H$ and $X_L$ as $X_H = \phi_H\left(1/\bar{f}\right)$ and $X_L = \phi_L\left(1/\bar{f}\right)$, where $\bar{f} = \max_{\theta,v} f_\theta(v)$. Note that $X_L > 0$ and $X_H > 0$. The Proposition then follows directly from Lemma 4 applied for $p_L^{\max} = p_H^{\max} = 1/\bar{f}$. Part (1): By part (1) of Lemma 4, the described contracts are incentive compatible. Moreover, they must be optimal because the firm cannot do better than induce first-best surplus and charge unconstrained optimal markups in both segments. Part (2): Follows directly from part (2) of Lemma 4.

**Proof of Proposition 11: (1) Sufficiently small**: Define $p_\theta^{\max} = \min\left\{1/\bar{f}, c/\left(1 - F_\theta(c)\right)\right\}$. Let $\tau > 0$ be sufficiently small such that (a) there is strict full-market-coverage and (b) $-\frac{1}{2}\phi_L(p_L^{\max}) \leq \tau(H - L) \leq \frac{1}{2}\phi_H(p_H^{\max})$, where $\phi_H(p_H^{\max})$ and $\phi_L(p_L^{\max})$ are defined above in equations (54)-(55).

**(2) Optimal markups:** For $\tau > 0$ sufficiently small, there is strict full-market-coverage in equilibrium and all customers strictly prefer to purchase from one of the two firms. Thus firm A's best response utility offer $U_\theta^A$ is always within an open interval for which residual demand from consumers of type $\theta$ is $G_\theta\left(U_\theta^A\right) = \frac{1}{2\tau_\theta}\left(U_\theta^A - U_\theta^B + \tau_\theta\right)$ and segment $\theta$ profit and its derivative are $\Pi_\theta^A\left(U_\theta^A\right) = \frac{1}{2\tau_\theta}\left(U_\theta^A - U_\theta^B + \tau_\theta\right)\left(S_\theta^A - U_\theta^A\right)$ and $\partial\Pi/\partial U_\theta^A = \frac{1}{2\tau_\theta}\left(-2U_\theta^A + S_\theta^A + U_\theta^B - \tau_\theta\right)$, respectively. Thus, defining $\tilde{U}_\theta^A = \arg\max_{U_\theta^A} \Pi_\theta^A\left(U_\theta^A\right)$ and $\tilde{\mu}_\theta^A = S_\theta^A - \tilde{U}_\theta^A$, we have $\tilde{U}_\theta^A = \frac{1}{2}\left(S_\theta^A + U_\theta^B - \tau_\theta\right)$ and $\tilde{\mu}_\theta^A = S_\theta^A - \tilde{U}_\theta^A = \frac{1}{2}\left(S_\theta^A - U_\theta^B + \tau_\theta\right)$. These are the profit maximizing utility offer and markup for segment $\theta$ treating $S_\theta^A$ as given while ignoring IC-L and IC-H. The latter can be re-written as $\tilde{\mu}_\theta^A = \frac{1}{2}\left(\mu_\theta^A + U_\theta^A - U_\theta^B + \tau_\theta\right)$. Given a symmetric equilibrium in which $U_\theta^A = U_\theta^B$ this simplifies as follows (dropping firm superscripts):

$$\tilde{\mu}_\theta = (\mu_\theta + \tau_\theta)/2. \tag{65}$$

(To clarify, $\mu_\theta$ is the actual markup chosen while $\tilde{\mu}_\theta$ would be the optimal markup to charge given $S_\theta$ if IC-L and IC-H were not a constraint. When IC constraints bind, the two may differ.)

**(3) Existence by construction:** Assume that each firm offers $v_{t\theta} = c$ and $\mu_\theta = \tau_\theta$. Then $-\phi_L(p_L^{\max}) < \mu_H - \mu_L < \phi_H(p_H^{\max})$ and hence Lemma 4 implies incentive compatibility. Moreover,

$\tilde{\mu}_\theta = \tau_\theta$ (by equation (65)). Thus contracts are best responses to each other as they implement first-best surplus and unconstrained optimal markups. Moreover, if $c/(1-F_L(c)) < 1/\bar{f}$ (and hence $c/(1-F_H(c)) < 1/\bar{f}$ by FOSD) then optimal penalty fees include $p_{3\theta} = p_\theta^{\max} = c/(1-F_L(c))$.

**(4) Uniqueness:** Start with a symmetric pure strategy equilibrium. Consider the following deviation from the equilibrium contract. Hold marginal prices constant but adjust markups to $\hat{\mu}_H = \tilde{\mu}_H - (\mu_H - \tilde{\mu}_H)$ and $\hat{\mu}_L = \tilde{\mu}_L - (\mu_L - \tilde{\mu}_L)$. Ignoring IC-L and IC-H for the moment, notice that this change in markups leaves profits unchanged. This follows because profits in segment $\theta$, $\Pi_\theta^A(\mu_\theta^A) = \frac{1}{2\tau_\theta}\left(S_\theta^A - U_\theta^B + \tau_\theta - \mu_\theta^A\right)\mu_\theta^A$, are a quadratic in $\mu_\theta^A$, and thus are symmetric in $\mu_\theta^A$ around the maximum at $\tilde{\mu}_\theta^A$. By definition of $\hat{\mu}_H$ and $\hat{\mu}_L$, the difference in revised markups is: $\hat{\mu}_H - \hat{\mu}_L = (2\tilde{\mu}_H - \mu_H) - (2\tilde{\mu}_L - \mu_L)$. Next, substituting in $\tilde{\mu}_\theta = (\mu_\theta + \tau_\theta)/2$ from equation (65), yields $\hat{\mu}_H - \hat{\mu}_L = \tau_H - \tau_L$.

Given the assumption that $\tau_H - \tau_L \in (0, \frac{1}{2}\phi_H(p_H^{\max})]$, Lemma 4 implies that at these markups first-best allocations are optimal: $v_{1L}^s = v_{2L}^s = v_{1H}^s = v_{2H}^s = c$ and all incentive constraints can be satisfied given the right choice of penalty fees. If allocations were not already first-best, this represents a strictly profitable deviation and a contradiction of the equilibrium contract being a best response. Thus the equilibrium contract must have first-best allocations: $v_{1L}^s = v_{2L}^s = v_{1H}^s = v_{2H}^s = c$. Moreover, as the difference $\hat{\mu}_H - \hat{\mu}_L$ is strictly inside the interval $[-\phi_L(p_L^{\max}), \phi_H(p_H^{\max})]$ for which Lemma 4 applies, to preclude a profitable deviation and contradiction, the markups $\hat{\mu}_H$ and $\hat{\mu}_L$ must be locally optimal ignoring IC-L and IC-H. That is $\partial\Pi/\partial\mu_L = \partial\Pi/\partial\mu_H = 0$ at first-best allocations. In this case the unique equilibrium is for markups to be $\mu_H = \tau_H$ and $\mu_L = \tau_L$.

### B.16 Proof of Corollary 2

**(1)** Total welfare result: With bill-shock regulation, equilibrium pricing matches the attentive case, and Proposition 9 implies that allocations are inefficient in all equilibria for any $\tau > 0$ (sufficiently small for strict full-market-coverage). In contrast, without bill-shock regulation, Proposition 11 shows that allocations are efficient for sufficiently small $\tau > 0$. Moreover, without bill-shock regulation, Proposition 11 shows equilibrium is symmetric so transportation costs are minimized. Thus bill-shock regulation strictly reduces welfare.

**(2)** Distributional result: Without bill-shock regulation, Proposition 11 implies IC-L and IC-H are slack, meaning $\frac{\partial\Pi}{\partial U_L} = -\frac{\partial\Pi}{\partial U_H} = 0$. With bill-shock regulation, Proposition 9 implies that in any symmetric equilibrium IC-H binds and $\frac{\partial\Pi}{\partial U_L} = -\frac{\partial\Pi}{\partial U_H} > 0$. Using superscript "BSR" to denote outcomes under bill-shock regulation, this implies high types win, $U_H^{BSR} > S_H^{FB} - \tau_H = \hat{U}_H$, but low types lose, $U_L^{BSR} < S_L^{BSR} - \tau_L < S_L^{FB} - \tau_L = \hat{U}_L$. Firms still split both segments equally, but now make less on high types $S_H^{FB} - U_H^{BSR} < \tau_H$ and more on low types $S_L^{BSR} - U_L^{BSR} > \tau_L$. On

average firms lose money. The first-order condition under bill-shock regulation ($\frac{\partial \Pi}{\partial U_L} = -\frac{\partial \Pi}{\partial U_H} > 0$) and symmetry ($G_\theta / g_\theta = \tau_\theta$) imply that

$$\frac{1}{2}\left(S_L^{BSR} - U_L^{BSR} - \tau_L\right)(1-\beta) = -\frac{\tau_L}{\tau_H}\frac{1}{2}\left(S_H^{FB} - U_H^{BSR} - \tau_H\right)\beta < -\frac{1}{2}\left(S_H^{FB} - U_H^{BSR} - \tau_H\right)\beta.$$

The inequality shows that the profit gain on low types is less than the profit loss on high types.

*Acknowledgments.* A previous version of this paper circulated under the title "Bill Shock: Inattention and Price-Posting Regulation". I thank Heski Bar-Isaac, Alessandro Bonatti, Glenn Ellison, Bob Gibbons, Ginger Jin, Bob Pindyck, and Tavneet Suri for careful reading and many helpful comments and suggestions. I also thank Igor Karagodsky for research assistance. Finally I am in debt to an anonymous referee for suggesting I model naively inattentive consumers.

## References

**Altschul, Michael F., Christopher Guttman-McCabe, and Brian M. Josef**, "In the Matter of Empowering Consumers to Avoid Bill Shock, CG Docket No. 10-207 and Consumer Information and Disclosure, CG Docket No. 09-158: Comments of CTIA - The Wireless Association," January 10, 2011. `http://fjallfoss.fcc.gov/ecfs/document/view?id=7021025497`.

**Armstrong, Mark and John Vickers**, "Competitive Price Discrimination," *RAND Journal of Economics*, Winter 2001, *32* (4), 579–605.

**_ and _** , "Consumer Protection and Contingent Charges," *Journal of Economic Literature*, 2012, *50* (2), 477–493.

**Aron-Dine, Aviva, Liran Einav, Amy Finkelstein, and Mark Cullen**, "Moral Hazard in Health Insurance: How important is forward looking behavior?," November 2012. `http://www.stanford.edu/~leinav/Forward_Looking.pdf`.

**Aumann, Robert J., Sergiu Hart, and Motty Perry**, "The Absent-Minded Driver," *Games and Economic Behavior*, 1997, *20* (1), 102–116.

**Banerjee, Abhijit V and Lawrence H Summers**, "On frequent flyer programs and other loyalty-inducing economic arrangements," September 1987. Harvard Institute of Economic Research Discussion Paper Number 1337. Available at `http://economics.mit.edu/files/501`.

**Battigalli, Pierpaolo**, "Dynamic Consistency and Imperfect Recall," *Games and Economic Behavior*, 1997, *20* (1), 31–50.

**Benoit, David**, "Customers Opt in for Overdraft Protection," *The Wall Street Journal*, November 26, 2010.

**Bernheim, B. Douglas and Antonio Rangel**, "Addiction and Cue-Triggered Decision Processes," *American Economic Review*, 2004, *94* (5), 1558–1590.

**Berry, Steven, James Levinsohn, and Ariel Pakes**, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841–890.

**Borenstein, Severin**, "To What Electricity Price Do Consumers Respond? Residential Demand Elasticity Under Increasing-Block Pricing," 2009. `http://faculty.haas.berkeley.edu/borenste/download/NBER_SI_2009.pdf`.

**Bubb, Ryan and Alex Kaufman**, "Consumer biases and mutual ownership," *Journal of Public Economics*, September 2013, *105*, 39–57.

**Bulkley, George**, "The role of loyalty discounts when consumers are uncertain of the value of repeat purchases," *International Journal of Industrial Organization*, 1992, *10* (1), 91–101.

**Cairns, Robert D. and John W. Galbraith**, "Artificial Compatibility, Barriers to Entry, and Frequent-Flyer Programs," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 1990, *23* (4), 807–816.

**Caminal, Ramon and Adina Claici**, "Are loyalty-rewarding pricing schemes anti-competitive?," *International Journal of Industrial Organization*, August 2007, *25* (4), 657–674.

**Courty, Pascal and Hao Li**, "Sequential Screening," *The Review of Economic Studies*, 2000, *67* (4), 697–717.

**Crémer, Jacques**, "On the Economics of Repeat Buying," *The RAND Journal of Economics*, 1984, *15* (3), 396–403.

**CTIA - The Wireless Association**, "CTIA-The Wireless Association, Federal Communications Commission and Consumers Union Announce Free Alerts to Help Consumers Avoid Unexpected Overage Charges," Press Release October 17, 2011. `http://www.ctia.org/resource-library/press-releases/archive/ctia-federal-communications-commission-consumers-union-announce-free-alerts`.

**Dash, Eric and Nelson D. Schwartz**, "Banks Seek to Keep Profits as New Oversight Rules Loom," *The New York Times*, July 15, 2010. `http://www.nytimes.com/2010/07/16/business/16wall.html`.

**DellaVigna, Stefano**, "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 2009, *47* (2), 315–372.

_ **and Ulrike Malmendier**, "Contract Design and Self-Control: Theory and Evidence," *The Quarterly Journal of Economics*, 2004, *119* (2), 353–402.

**Deloney, Amalia, Linda Sherry, Susan Grant, Parul P. Desai, Chris M. Riley, Matthew F. Wood, John D. Breyault, Jessica J. Gonzalez, and Benjamin Lennett**, "In the Matter of Empowering Consumers to Avoid Bill Shock, CG Docket No. 10-207 and Consumer Information and Disclosure, CG Docket No. 09-158: Comments of the Center for Media Justice, Consumer Action, Consumer Federation of America, Consumers Union, Free Press, Media Access Project, National Consumers League, National Hispanic Media Coalition and New America Foundation Open Technology Initiative in response to notice of proposed rulemaking," January 10, 2011. `http://fjallfoss.fcc.gov/ecfs/document/view?id=7021025418`.

**Department of Justice**, "Justice Department Files Antitrust Lawsuit to Block AT&Ts Acquisition of T-Mobile," Office of Public Affairs Press Release August 31, 2011. `http://www.justice.gov/opa/pr/2011/August/11-at-1118.html`.

**Diamond, Peter A.**, "A model of price adjustment," *Journal of Economic Theory*, 1971, *3* (2), 156–168.

**Eliaz, Kfir and Ran Spiegler**, "Contracting with Diversely Naive Agents," *The Review of Economic Studies*, 2006, *73* (3), 689–714.

_ **and** _ , "Consumer Optimism and Price Discrimination," *Theoretical Economics*, 2008, *3* (4), 459–497.

**Ellison, Glenn**, "A Model of Add-on Pricing," *The Quarterly Journal of Economics*, March 2005, *120* (2), 585–637.

**Farrell, Joseph and Paul Klemperer**, "Coordination and Lock-in: Competition with Switching Costs and Network Effects," in Mark Armstrong and Robert H. Porter, eds., *Handbook of Industrial Organization*, Vol. 3, Elsevier, 2007, pp. 1967–2072. Chapter 31.

**Federal Reserve Board**, "Regulation E final rule," *Federal Register*, November 17, 2009, *74* (220), 59033–59056.

**Gabaix, Xavier and David Laibson**, "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, May 2006, *121* (2), 505–540.

**Genachowski, Julius**, "Prepared Remarks of Chairman Julius Genachowski," FCC October 13, 2010. http://reboot.fcc.gov/fcc-s-consumer-empowerment-agenda.

**Grubb, Michael D.**, "Selling to Overconfident Consumers," *American Economic Review*, December 2009, *99* (5), 1770–1807.

_ , "Consumer Inattention and Bill-Shock Regulation," July 5 2012. MIT Sloan Research Paper No. 4987-12. Available at SSRN:http://dx.doi.org/10.2139/ssrn.1983518.

_ **and Matthew Osborne**, "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock," *The American Economic Review*, forthcoming.

**Heidhues, Paul, Botond Kőszegi, and Takeshi Murooka**, "Exploitative Innovation," 2012. http://www.personal.ceu.hu/staff/Botond_Koszegi/exploitative_innovation.pdf.

_ **,** _ **, and** _ , "Inferior Products and Profitable Deception," 2014. http://www.personal.ceu.hu/staff/Botond_Koszegi/inferior_products.pdf.

**Inderst, Roman and Marco Ottaviani**, "Sales Talk, Cancellation Terms and the Role of Consumer Protection," *The Review of Economic Studies*, July 2013, *80* (3), 1002–1026.

**Ito, Koichiro**, "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing," *The American Economic Review*, 2014, *104* (2), 537–563.

**Jiang, Lai**, "The Welfare Effects of 'Bill Shock' Regulation in Mobile Telecommunication Markets," November 2013. http://www.lai-jiang.com/uploads/Jiang_JobMarketPaper.pdf.

**Jolson, Marvin A., Joshua L. Wiener, and Richard B. Rosecky**, "Correlates of Rebate Proneness," *Journal of Advertising Research*, 1987, *27* (1), 33–43.

**Khouja, Moutaz**, "A joint optimal pricing, rebate value, and lot sizing model," *European Journal of Operational Research*, 10/16/ 2006, *174* (2), 706–723.

**Laibson, David**, "Golden Eggs and Hyperbolic Discounting," *The Quarterly Journal of Economics*, 1997, *112* (2), 443–477.

**Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips**, "Calibration of Probabilities: The State of the Art to 1980," in Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, 1982, pp. 306–334.

**Liebman, Jeffrey B. and Richard Zeckhauser**, "Schmeduling," October 2004. http://www.hks.harvard.edu/fs/rzeckhau/Schmeduling_Oct172004.pdf.

**Martin, Andrew**, "Bank of America to End Debit Overdraft Fees," *The New York Times*, March 10, 2010. http://www.nytimes.com/2010/03/10/your-money/credit-and-debit-cards/10overdraft.html.

**Miao, Chun-Hui**, "Consumer myopia, standardization and aftermarket monopolization," *European Economic Review*, 2010, *54* (7), 931–946.

**Pavan, Alessandro, Ilya Segal, and Juuso Toikka**, "Dynamic Mechanism Design: A Myersonian Approach," *Econometrica*, March 2014, *82* (2).

**Piccione, Michele and Ariel Rubinstein**, "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior*, 1997, *20* (1), 3–24.

**Rochet, Jean-Charles and Lars A. Stole**, "Nonlinear Pricing with Random Participation," *The Review of Economic Studies*, 2002, *69* (1), 277–311.

**Saez, Emmanuel**, "Do Taxpayers Bunch at Kink Points?," June 2002. `http://eml.berkeley.edu/~saez/bunch.pdf`.

**Sims, Christopher A.**, "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

_ , "Chapter 4 - Rational Inattention and Monetary Economics," in Benjamin M. Friedman and Michael Woodford, eds., *Handbook of Monetary Economics*, Vol. 3, Elsevier, 2010, pp. 155–181.

**Soman, Dilip**, "The Illusion of Delayed Incentives: Evaluating Future Effort-Money Transactions," *Journal of Marketing Research*, 1998, *35* (4), 427–437.

**Spiegler, Ran**, *Bounded Rationality and Industrial Organization*, New York: Oxford University Press, 2011.

**Stango, Victor and Jonathan Zinman**, "What do Consumers Really Pay on Their Checking and Credit Card Accounts? Explicit, Implicit, and Avoidable Costs," *American Economic Review Papers and Proceedings*, 2009, *99* (2).

_ **and** _ , "Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Bank Overdraft Fees," *Review of Financial Studies*, 2014, *27* (4), 990–1030.

**Yao, Song, Carl F. Mela, Jeongwen Chiang, and Yuxin Chen**, "Determining Consumers' Discount Rates with Field Studies," *Journal of Marketing Research*, December 2012, *49* (6), 822–841.