Title
Essays on Quantitative Marketing Theory

Author
Zhou, Zihao

Publication Date
2021

Peer reviewed|Thesis/dissertation

Essays on Quantitative Marketing Theory

By

Zihao Zhou

A dissertation submitted in partial

satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Yuichiro Kamada, Co-Chair

Professor J. Miguel Villas-Boas, Co-Chair

Associate Professor Haluk Eergin

Associate Professor Zsolt Katona

Spring 2021

Essays on Quantitative Marketing Theory

by

Zihao Zhou

Abstract

Essays on Quantitative Marketing Theory

By

Zihao Zhou

Doctor of Philosophy in Business Administration

University of California, Berkeley

Associate Professor Yuichiro Kamada, Co-Chair

Professor J. Miguel Villas-Boas, Co-Chair

In three essays, I present my work that uses mathematical modelling to analyse real-world problems in marketing. In the first chapter, I model the design of referral programmes and offer normative advice about improving the cost-effectiveness of a referral programme. The second chapter is joint work with my co-advisor, Yuichiro Kamada. In this work, we construct a tractable game-theoretic model for multiple-priority queue management. Lastly, the third chapter analyses the incentives of online bloggers that post product reviews to subscribers when the reviews may be sponsored.

Customer referrals have become an increasingly important way for firms to grow their customer base. Likely thanks to improvements to customer relationship management software as well as growing awareness of the potential of word-of-mouth marketing, referral programmes are becoming increasingly popular among firms. In the first chapter, I study the design of customer referral programmes by constructing a stylized static principal-agent model with hidden actions, in which a firm designs a referral programme to incentivize an existing customer to exert costly efforts to refer the customer's friends to the firm. In the baseline model, I find it optimal for the firm to pay the customer if and only if every friend of the customer is successfully referred. In a number of extensions that are important and relevant to referral programmes, although the optimal referral contract is no longer a threshold contract, this class of contracts still plays an important role in the optimal referral programme design. Overall, my work shows that it is cost-effective to use or include threshold contracts to incentivize efforts.

When access to a service facility is congested, service providers commonly implement a special type of queue called priority queue, where each person/entity in the queue has an associated priority such that those with a higher priority will be ahead of those with a lower priority in the queue. Previous studies into second-degree price discrimination and queue management suggest that the firm that manages the firm that manages the park should set a large number of different priorities in order to improve price discrimination. In the second chapter, my co-author and I construct a model in which an amusement park sells different priority passes to customers in a queue whose utilities depend on positions in the queue. A customer's valuation of a priority pass depends on the distribution of customers buying each priority pass, and hence other customers' pur-

chase decisions have an externality on the customer's valuation, which differentiates our model from the standard screening models. Through the model, we discuss the implementability of selling multiple passes for different patterns of customer utility functions. The main result of our work is that the externality makes the implementation of multi-pass schemes difficult, an issue that persists even when customers have heterogeneous utilities of positions in a queue.

Thanks to the popularity of internet platforms such as YouTube and TikTok, more and more customers are turning to their favourite internet bloggers for product reviews. That the bloggers have many subscribers that rely on their reviews for purchase decision makes bloggers a potential sales channel. This type of marketing is often called influencer marketing. To incentivize the bloggers to help promote a firm's product, one common incentive is for the firm to sponsor a blogger's product review by offering sales commissions for purchases by the blogger's subscribers. In a sponsored review, since the blogger now benefits from higher sales, the blogger has a bigger incentive to review the product favourably, even though the blogger's own private signal about the product's quality says otherwise. On the other hand, the blogger cares about the accuracy of the review because accurate reviews attract viewership. In the third chapter, I construct a stylized model that analyses the incentives of a blogger posting reviews of products with uncertain quality on an internet platform to the blogger's subscribers. Specifically, the blogger receives a costless private signal about the quality of the product and then sends a review message to the subscribers. On the one hand, the platform rewards the blogger for posting accurate reviews. On the other hand, higher sales of the product lead to higher sales commission offered by the firm manufacturing the product. I find that the blogger has an incentive to truthfully communicate the received signal when the signal is informative; otherwise, the blogger has an incentive to pretend to receive a favourable signal even when the realized signal is unfavourable to the product. Under a symmetric signal structure, my model shows that the blogger has an incentive for an honest review when the product is perceived to be mid-end.

Acknowledgements

I am indebted to my advisors, Yuichiro Kamada and J. Miguel Villas-Boas, for their generous mentorship and encouragement throughout my doctoral study. I am grateful to my committee members, Haluk Ergin and Zsolt Katona, for their guidance for my dissertation. I would also like to thank Ganesh Iyer and Przemyslaw Jeziorski for their help with my academic endeavour. This dissertation has received a lot of great comments and feedback from many other faculty and students at and outside UC Berkeley. All errors remain with me.

Completing a PhD programme is more than just academic endeavours. I would like to thank Wei Lin, Ao Wang, and Chaoran Yu for the great dining experience in the Bay Area. As someone that has procrastinated over and over again on getting my driving licence, I cannot possibly have such a nice travel experience around the Bay Area without Jesse Yao as a great travel companion and reliable driver. I would like to thank Vincent Skiera and Yipei Zhang, together with many friends already mentioned here, for many board-game nights that have reminded me that life is more than just academics. I would like to thank Yunbo Liu for being a kind and considerate *kouhai*, as well as a great board-game player. I am also grateful to Xiaohan Yan and Shuo Zhang for many interesting discussions about mathematics.

Lastly, I would like to thank my family and friends that have already been supportive of me. Your support has definitely carried me through the doctoral study.

To all that are coping or struggling with stammer,

Contents

Chapter 1

On the Design of Referral Programmes

1.1 Introduction

Customer referrals have become an increasingly important way for firms to grow their customer base. Kumar, Petersen, and Leone (2007) find that customers that refer new customers are the most valuable; through field and laboratory experiments, results of Garnefeld et al. (2013) suggest that referral programmes improve customer loyalty; Schmitt, Skiera, and Van den Bulte (2011) show that referred customers tend to have a higher contribution margin as well as a higher retention rate. Likely thanks to improvements to customer relation management software as well as growing awareness of the potential of word-of-mouth marketing, referral programmes are becoming increasingly popular among firms.

While exact implementations vary considerably, every referral programme can be considered as a scheme in which a firm's existing customer gets rewarded (monetarily or non-monetarily) for referring new customers to the firm.[1] For example, Dropbox, a large cloud file-hosting service provider, rewards customers with more free storage for making new referrals. Helping the company grow its customer base from about $100,000$ to four million in 15 months, Dropbox's referral programme has been widely touted as one of the best examples of a successful referral programme.

Another example of referral programme comes from Pinduoduo, which is the second-largest e-commerce company in China. The company has been incentivizing its customers to refer their friends to the platform with discounts soon after its launch in 2015. Specifically, the company promises to sell a product to a customer for free if that customer manages to refer enough friends to the platform within 24 hours, and the customer would get nothing if the number of referrals falls short of the threshold. Thanks to its referral programme, Pinduoduo was able to grow its user base from only 10 million in 2016 to more than 600 million in 2020.

Designing a referral programme has many dimensions to consider, such as the externality between an existing/focal customer and the customer's friends, the number of friends an existing customer has, and how hard the existing customer exerts efforts towards successful referrals. In

[1] In this paper, referral rewards, either monetary or non-monetary, are encoded by the utility of the customer. See Jin and Huang (2014) for a comparison on the effects of monetary and non-monetary referral rewards.

light of the complexities of designing a referral programme, I qualify the scope of this paper and focus on the hidden action part of designing a referral programme: a customer can choose how large efforts to put in the referral attempts. In other words, the customer's private efforts are assumed to have an impact on the conversion rate of the customer's friends: preparing a detailed and comprehensive summary such as a short demonstration of the customer's own user experience to the customer's friends should tend to have a higher conversion rate than just sending an uninformative referral link.

Overall, this paper finds it cost-effective to use or include threshold contracts to incentivize efforts. A threshold contract is a step function with respect to the number of referrals. One example of an actual referral programme that uses a threshold contract is the one used by Pinduoduo described earlier. The use of threshold contracts in referral programmes is popular among e-commerce companies in China that need market share growth. For example, many companies in the online grocery shopping sector, which is rapidly growing in China, adopt threshold-based referral programmes.

The main result of the baseline model is that it is optimal for the firm to pay the customer if and only if every friend of the customer's is converted. In the meantime, the baseline model provides a complete cost-effectiveness ordering among different threshold contracts with very interpretable insights. While using only threshold contracts is not always optimal in a number of important extensions that are relevant to referral programmes, the paper shows that the inclusion of threshold contracts still plays an important part in incentivizing efforts.

To motivate this paper's model, consider the following minimal binary-outcome example. Suppose that it is common knowledge that an existing customer has a friend that is a potential new customer to the firm. The firm can design a referral programme in which the customer gets paid for bringing in the new customer. Specifically, a referral contract specifies the referral rewards the customer gets when the friend is converted and when the friend is not. In this case, the customer's optimal effort level is characterized by the customer's first-order condition, and the firm just needs to find the profit-maximizing effort level subject to the customer's first-order condition.

When the number of friends that are potential new customers is more than one, however, the analysis is not straightforward. Suppose that with a fixed effort level, each of the customer's friend becomes the firm's new customer independently with the same probability. When the customer has multiple friends, the firm's choice set of contracts is a multi-dimensional vector space, which can be very large. Under this more general setup, the patterns of optimal contracts are unclear. One form of contracts to consider is the linear contracts, in which the customer gets a fixed additional reward for each additional referral. Focusing on linear contracts is straightforward: in a linear contract, it can be shown that the customer's marginal rewards is constant in efforts, and thus the customer's optimal effort level is characterized by the customer's first-order condition. The firm then just needs to solve for the profit-maximizing effort level subject to the customer's first-order condition. Despite the popularity and ease of analysis, with little known about the cost-effectiveness of the much larger class of non-linear contracts, linear contracts could be suboptimal, which this paper's result of highest-threshold optimality implies.

In existing literature on referral programmes, the existing customer usually has a binary effort space: the customer decides whether to send a referral message to a friend at a fixed cost, and the friend gets converted with a fixed probability. In this paper, efforts affect the distribution of referral outcomes continuously. Furthermore, most studies focus on the referral of just one friend, whereas this paper allows for the conversions of multiple friends, which greatly expand the consideration set of referral contracts for the firm.

Since the firm cannot observe the customer's chosen effort level, this paper's baseline is a hidden-action model with finite outcomes. There are a few assumptions in the baseline model. The first is that the customer's number of referrable friends is common knowledge. An empirical interpretation of this assumption is that the firm has a good picture of the customer's social network as well as the convertibility of the customer's friends. This assumption, though unlikely to be entirely accurate, is highly relevant in today's world of big data and hence the assumption's implications are worth analysing. In an extension of the baseline model, the number of referrable friends becomes private information of the customer, and the result of highest-threshold optimality will be shown to hold for the customer with the larger number of referrable friends.

Another assumption is that if the customer chooses effort level q, then every friend becomes a new customer independently with probability q. Under the context of referral programmes, this assumption means that the customer's efforts spill over to the customer's friends' conversion homogeneously and independently. One interpretation is that the customer's referral efforts are freely applicable to each friend's conversion. For example, suppose that a customer is to recommend the firm's new gadget to the customer's friends. After spending time and efforts looking into the specifications of the gadget and preparing the summary of the customer's own user experience, the customer can present the customer's referral preparation to the friends at no extra cost. This assumption may not hold precisely hold in practice, i.e., different friends may respond to the same referral efforts differently. Still, given that referral efforts for one friend are likely to be at least partially applicable towards the referral attempt for another friend, this assumption is a good starting point to analyse.

Lastly, the customer is assumed to have limited liability in the sense whatever the realized referral outcome is, the firm cannot charge the customer. This assumption is reasonable under the context of referral programmes as there are few actual referral programmes where the customer may end up owing the firm money.

In the next section, I discuss the existing literature relevant to this paper in terms of the modelling approach and the topic. Section 1.3 introduces the baseline model, followed by extensions of the model in Section 1.4. Section 1.5 analyses the case when the number of referral friend is private information of the existing customer. Lastly, to make the baseline model more relatable to studies into the broader principal-agent problem, Section 1.6 provides and discusses sufficient conditions under which the result of highest-threshold optimality still holds.

1.2 Literature Review

There has been substantial work on the study of the principal-agent problem with hidden actions. Grossman and Hart (1983) provide a general model for this class of problems with finite outcomes. This paper has more restrictions that are relevant to referral programmes. These new assumptions lead to sharper results about the patterns of optimal referral contracts. To be specific, the binomial distribution, the first-order approach,[2] and the limited liability of customer assumption make the highest-threshold contracts the most cost-effective form of contract.

This paper is also closely related to the study of salesforce management. Basu et al. (1985) provide a general theoretical framework in which a firm designs a quantity-dependent contract to induce a salesperson's private efforts. That paper discusses the comparative statics of different structural parameters on the shape of the optimal salesforce compensation plan. In a later paper, Raju and V. Srinivasan (1996) restrict the attention to the class of quota-based compensation plan, which is analogous to the class of threshold contracts in this paper. Through numerical experiments, the authors show that the use of quota-based contracts leads to only small suboptimality in comparison to the general optimal compensation scheme in Basu et al. (1985). In contrast to these two papers, this paper has different functional assumptions that lead to sharper and more interpretable characterizations of the firm's optimal contract in a number of settings that are more relevant to referral programmes than to salesforce management.

The result in the baseline model that the highest threshold contracts are optimal is also established in Balmaceda et al. (2016) with a similar mathematical setup; Innes (1990) and Poblete and Spulber (2012) have results of similar flavour in the sense that these two papers with their applications find it optimal to concentrate payments in extreme outcomes. In contrast to these papers with similar highest-threshold-optimality results, the baseline model in this paper provides more interpretable insights about the cost-effectiveness of using high-threshold contracts, which are useful for managerial recommendation. Additionally,in doing so, the baseline model also provides a complete ordering of different threshold contracts based on cost-effectiveness. Moreover, this paper enriches the baseline model in a number of extensions that are highly relevant and important to referral programmes. In these extensions, threshold contracts are not always optimal, but the paper still shows that the inclusion of threshold contracts still plays an important role in the optimal referral contract.

This paper is not the first in which the customer can have multiple friends; nor is it the first to look into non-linear referral contracts. Lobel, Sadler, and Varshney (2017) construct a dynamic model to study the design of referral programmes. The authors find that non-linear contracts are useful in screening the number of referrable friends an existing customer has, whereas linear contracts can be used to compensate customers for their effort costs. Incidentally, this paper finds linear contracts good for compensating efforts, whereas threshold contracts are great at incentivizing efforts. With respect to focus, Lobel, Sadler, and Varshney (2017) focus on private information and simplifies the effort-incentivizing consideration, whereas this paper provides a richer analysis into the effort incentivizing through a referral programme. Specifically with respect to the modelling of hidden action, in Lobel, Sadler, and Varshney (2017), efforts are applicable to only one friend, whereas this paper assumes efforts can be freely reused towards the conversion of other friends.

As for the topic of the referral programme, there are several theoretical studies with different focuses. Kamada and Öry (2020) study whether referral programmes are complements or substitutes for free contracts in the presence of externalities between the existing customer and the

[2]There are several moral-hazard models using the first-order approach, such as Hölmstrom (1979) and Mirrlees (1976). See Rogerson (1985) and Jewitt (1988) for some excellent discussions on sufficient conditions that make the first-order approach valid.

customer's friend. The authors find that the two are complements when externalities are small and substitutes when the externalities are significant.

Biyalogorsky, Gerstner, and Libai (2001) have a similar goal to that of Kamada and Öry (2020), yet with a different modelling approach. That paper assumes that the customer can be made "delighted" through direct price cuts and treats referral rewards as indirect price cuts that are useful in preventing free-riding on the discounts. The paper shows that referral rewards become indispensable to the firm's profits when it is difficult to make the customer delighted through direct price cuts.

Lastly, Leduc, Jackson, and Johari (2017) use a dynamic social learning model in which connected agents infer from their neighbours' adoption choices of the firm's product to make their own adoption decisions. That paper's main result is that referral rewards can be better than inter-temporal price discrimination in some network structures.

1.3 Baseline Model

Assume there is a firm and an existing/focal customer. I build a static model through which the firm incentivizes the customer to refer their friends to the firm. Assume that it is common knowledge that the customer has N friends/leads that are potential new customers. We call these friends referrable friends because their conversion rates can be influenced by the existing customer's referral efforts. Assume each customer has a fixed customer lifetime value; in other words, the firm gets a fixed profit from each new customer, denoted by $\pi > 0$.[3]

The customer can exert effort level $q \in [0, 1]$ with disutility $c(q)$ such that each of the customer's N friends is going to adopt the firm's product with probability q independently. Assume the cost function $c(q)$ is continuously twice differentiable with $c(0) = 0$, $c'(q) > 0$, and $c''(q) \geq 0$ for all $q \in (0, 1)$. The actual number of converted friends is a non-negative integer-valued random variable ranging from 0 to N, which is denoted by $M(q, N)$. By setup, $M(q, N)$ is a random variable following the binomial distribution with N trials and success rate q. Let $F(\cdot; q, N)$ be the cumulative density function induced by effort level q, with $f(\cdot; q, N)$ being its probability mass function for $0 \leq n \leq N$.[4] One economic interpretation of the effect of efforts on the number of referrals is that efforts have an independent and stochastically homogeneous effect on the conversions of the customer's N friends.

1.3.1 First Best

First look at the first-best case where q is contractible. In this case, the referral contract between the firm and the customer can be of the form $(\beta, q) \in \mathbb{R} \times [0, 1]$, in which the firm pays the customer reward β and the customer chooses effort level q. Since q is contracted, the customer accepts the

[3]One extended interpretation of π, which makes the model more powerful, is that π is the ex-ante continuation payoff the firm gets from each new referral in a dynamic setting where the referred friends can continue to refer more friends to the firm.

[4]N as a variable of a function is omitted when there is no ambiguity, such as when N does not change in a given context.

contract as long as the customer's expected utility is above reserve utility, which is normalized to 0. Therefore, the firm solves the following constrained optimization problem:

$$\max_{\beta\in\mathbb{R},q\in[0,1]} \quad \pi qN - \beta, \quad \text{s.t.} \quad \beta - c(q) \geq 0. \tag{1.1}$$

The solution is unique and the constraint binds at the solution. Let $q_{FB}(N)$ denote the effort level in the solution. If $q_{FB} < 1$, $q_{FB}(N)$ is implied from the customer's first-order condition $c'(q_{FB}(N)) = \pi N$. For the rest of this paper, assume that in the solution to (1.1), the first-order condition holds.

1.3.2 Second Best

Now assume that the firm cannot observe the effort level chosen by the customer. Thus the firm and the customer cannot sign a contract on efforts. However, since the number of referrals is contractible and the customer's efforts affect its distribution, the firm can sign a contract with the customer on the realized value of $M(q, N)$, i.e., the realized number of referrals.

A **referral contract** is an $(N + 1)$-dimensional non-negative real vector $b = (b_0, b_1, \dots, b_N)$, with b_n being the rewards the firm gives the customer when $M(q, N) = n$. Given a referral contract b, let $v(b; q, N) = E\left[b_{M(q,N)}\right]$ be the expected referral rewards the customer gets from the referral contract. By the assumptions so far, $v(b; q, N)$ is twice continuously differentiable in b and q. Assume the utility of the customer's outside option to be 0.[5] The customer chooses effort level $q \in [0, 1]$ to maximize the customer's expected utility $v(b; q, N) - c(q)$. By the continuity of $f(n; q)$ and $c(q)$ in $q \in [0, 1]$ for each n, an optimal effort level exists. Given a contract b and effort level q, let $U(b; q, N) = \pi Nq - v(b; q)$ be the firm's profit if a customer with N referrable friends chooses effort level q in b. Taking into account of the customer's response to each referral contract, the firm designs its referral programme to maximize the firm's expected profits:

$$\max_{b\in\mathbb{R}^{N+1},q\in[0,1]} \quad U(b; q) = \pi Nq - v(b; q) \tag{1.2}$$

$$\text{s.t.} \quad v(b; q) \geq c(q), \tag{1.2a}$$

$$v(b; q) - c(q) \geq v(b; q') - c(q'), \quad \forall q' \in [0, 1], \tag{1.2b}$$

$$b_n \geq 0, \quad 0 \leq n \leq N, \tag{1.2c}$$

where (1.2a) is the customer's **participation constraint** and (1.2b) denotes the continuum of **effort-choice constraints**[6] that the customer must find the effort level optimal. Assumption (1.2c) is the limited liability of customer assumption that the firm can never charge a customer in a referral programme, which is commonly observed in actual referral programmes. Given the optimization problem (1.2), (b, q) in (1.2) is said to be **feasible** if (b, q) satisfies all the constraints

[5] This assumption is non-trivial: by setting the reserve utility to 0, every non-negative contract will satisfy the customer's participation constraint, which plays a role in the existence of second-best contracts. See Grossman and Hart (1983) and Innes (1990) for the role of the reserve utility.

[6] The definition of incentive compatibility is reserved for the case where the number of referrable friends is private information.

of (1.2). Given a contract b, the contract is said to be **strictly profitable** with respect to (1.2) if there exists some q such that (b, q) is feasible in (1.2) and $U(b; q) > 0$. Contract b is said to be **optimal** if there exists q such that (b, q) solves (1.2).[7]

Given an effort level θ and a referral contract b, the contract b is said to **induce** θ if θ is the unique optimal effort level to the customer in b. Graphically, the stronger effort-inducing definition ensures that $v_q(b; q)$ crosses $c'(q)$ only once on $(0, 1)$ at θ and from above.

The first-order approach is used to simplify the customer's effort-choice constraints, i.e., the continuum of effort-choice constraints in (1.2b) is simplified to one single constraint with respect to the customer's first order condition. Given a contract and effort level, the first-order condition is valid if the contract induces this effort level. Mirrlees (1999) shows that in general the first-order approach is invalid: provided that effort level at the firm's optimum is interior, the constraint substitution enlarges the firm's feasible set and hence the solution to the firm's relaxed optimization could differ from the actual problem. Rogerson (1985) shows that with two assumptions, MLRP[8] and the convexity of the distribution function condition (CDFC),[9] the first-order approach is valid. Conditional on that the first-order approach is valid, it can be shown that every profit-optimal contract must be non-decreasing, i.e., $b_0 \leq b_1 \leq \cdots \leq b_N$. It turns out that the effort-induced binomial distribution has MLRP, but not CDFC.[10]

It can be shown that the first-order approach is valid for the class of N-threshold contract if the increasing marginal cost of probability (IMCP) condition defined in Balmaceda et al. (2016) holds, which is a single-crossing property between the marginal cost function and the marginal cumulative distribution function with respect to efforts.[11]

Definition 1 (IMCP)**.** IMCP holds if $\frac{c'(q)}{f_q(N;q)}$ is increasing in $q \in (0, 1)$.

One example of cost function that together with the binomial distributional assumption satisfies IMCP can be found from the power cost function of the form $c(q) = \gamma q^M$ for some $M > N$ and large $\gamma > 0$. For the rest of the paper, assume IMCP holds.

Given a contract $b \in \mathbb{R}^{N+1}$ and $0 < n \leq N$, contract b is said to be an **n-threshold contract** if $0 = b_0 = b_1 = \cdots = b_{n-1} < b_n = b_{n+1} = \cdots = b_N$. Since IMCP makes the first-order approach valid for the class of N-threshold contracts, for each interior effort level, there exists an N-threshold contract that induces the effort level. In fact, it turns out that whenever an N-threshold contract induces some effort level, for every $1 \leq n \leq N$, there exists an n-threshold contract inducing the same effort level.

Proposition 1 (Threshold effort-inducing)**.** *Fix $\theta \in (0, 1)$. For each $1 \leq n \leq N$, if there exists an N-threshold contract inducing θ, then there exists an n-threshold contract inducing θ.*

[7] The definition of optimal contract or menu will change analogously with respect to the change of the firm's optimization problem.

[8] Given that $f(n; q) > 0$ for $q \in (0, 1)$, the effort-induced distribution is said to satisfy MLRP if $\frac{f_q(n;q)}{f(n;q)}$ is non-decreasing in n for $q \in (0, 1)$.

[9] Let $h(\cdot)$ be the inverse function of the cost function $c(\cdot)$, which is well-defined since $c'(q) > 0$ for every $q \in (0, 1)$. The effort-induced distribution is said to satisfy CDFC if $F(n; h(x))$ is convex for every x in $c([0, 1])$.

[10] In Appendix 1.A.1, CDFC is shown not to hold with any twice differentiable convex $c(\cdot)$.

[11] Appendix 1.A.2 provides the result that for every target interior effort level, there exists some N-threshold contract that that induces the effort level.

Proof. Fix $\theta \in (0,1)$. For every $1 \leq n \leq N$, given an n-threshold contract $t(n) \in \mathbb{R}^{N+1}$, $v_q(t(n);q) = -t_n(n)F_q(n-1;q) > 0$ since $F_q(n-1;q) < 0$ by the binomial distribution assumption. Therefore, for every $1 \leq n \leq N$, there exists an n-threshold contract such that $v_q(t(n);\theta) = c'(\theta)$. It remains to show that $t(n)$ induces θ.

Let $t(N) \in \mathbb{R}^{N+1}$ be the N-threshold contract inducing θ. Since the customer's first-order conditions hold at θ in both contracts, $v_q(t(n);\theta) \geq v_q(t(N);\theta)$ and hence

$$\binom{N}{n} q^{n-1}(1-q)^{N-n} \geq Nq^{N-1} \iff \left(\frac{q}{1-q}\right) \leq \frac{\binom{N}{n}}{N}.$$

The inequality on the right-hand side holds for $q = \theta$ and strictly holds for $q \in (0,\theta)$. Thus $v_q(t(n);q) > v_q(t(N);q)$ for $q \in (0,\theta)$ and $v_q(t(n);q) < v_q(t(N);q)$ for $q \in (\theta,1)$. Therefore, $v_q(t(n);q)$ crosses $c'(q)$ exactly once on $(0,1)$ at θ from above, and hence $t(n)$ induces θ. □

It turns out that given the existence of a strictly profitable contract, every optimal contract is necessary an N-threshold contract, i.e., a highest-threshold contract.

Proposition 2 (Threshold optimality). *An optimal contract exists, i.e.,* (1.2) *has a solution. Assume IMCP holds. If there exists a contract that is strictly profitable with respect to* (1.2)*, then every optimal contract is an N-threshold contract, with the customer choosing a strictly positive effort level.*

The proof first shows that an optimal contract in which the customer exerts a strictly positive effort level exist, and then shows that it is uniquely cost-efficient to use an N-threshold contract to incentivize that effort level.

Remark 1. It can be shown that the effort level chosen in an optimal contract is strictly less than $q_{FB}(N)$, the first-best effort level. Indeed, if $\theta > 0$ is the effort level induced in an optimal N-threshold contract, then $\pi N \geq v_q(t;\theta) = c'(\theta)$ for the firm's optimality condition. Let $q(t)$ be the optimal effort level the customer chooses in an N-threshold contract $t \in \mathbb{R}^{N+1}$. The effort level $q(t)$ is increasing in t_N, which is the referral reward from contract t when the outcome is N. Thus at the optimal contract t, $\frac{dv(t;q(t))}{dq} > v_q(t;q(t))$ when $q(t) = \theta$, where the right-hand side is the marginal rewards to the customer in contract t at effort level $q(t)$. Hence $\theta < q_{FB}$ by the strict convexity of the effort cost function. In contrast, Grossman and Hart (1983) show that the principal (adapted to this paper) can obtain the first-best payoff by setting $b_n = \pi n - \pi N q_{FB} - c'(q_{FB})$ for $0 \leq n \leq N$. Under this contract, the principal would obtain the first-best payoff in each outcome, and the agent would find it optimal to choose q_{FB} at which the customer's first-order condition holds and the participation constraint binds. This contract, however, is not allowed here, because in this contract, the customer needs to pay when $M(q) = 0$, which violates the customer's limited liability constraint.

To explain the intuition of the cost-effectiveness of highest-threshold contracts, I make the simplifying assumption that the firm is committed to a non-decreasing referral contract, i.e., $b_0 \leq b_1 \leq \cdots \leq b_N$. The assumption is not necessary for proving Proposition 2 but it is reasonable for referrable programmes and useful in providing interpretable insights. When we focus on non-decreasing contracts, it suffices to focus on the class of threshold contracts because it turns out that every non-decreasing contract is a convex combination of different threshold contracts inducing the same effort level.

Proposition 3 (Non-Decreasing contracts as convex hull of threshold contracts)**.** *Fix $\theta \in (0,1)$. Let $t(n)$ be the n-threshold contract inducing θ. If b is a non-decreasing contract such that $b_0 = 0$ and $v_q(b;\theta) = c'(\theta)$, then b induces θ and is in the convex hull of $\{t(n)\}_{n=1}^{N}$.*

Since the set of threshold contracts spans the set of non-decreasing contracts, it is sufficient to explain how the highest-threshold contracts are the most cost-effective among all threshold contracts. To do so, it is instructive to explain how threshold contracts can be cost-effective as second-best approximation of first-best contracts in which efforts are contractible. In the hypothetical first-best world where efforts are contractible, if the firm wishes to incentivize certain effort level, the contract can be a step function of efforts so that referral reward is zero for sure when the chosen effort level is below the contract requirement. In contrast, what happens in a threshold contract is that the customer is likely to get nothing if the chosen effort level is too low below the target effort level. In other words, a threshold contract can be interpreted as a second-best attempt to mimic the effect of a first-best contract that directly contracts on efforts.

I now explain why the highest-threshold contracts are the most cost-effective among all threshold contracts. It turns out that for every arbitrary target effort level, it is always more cost-effective to use a higher-threshold contract than to use a lower-threshold contract to incentivize that effort level.

Proposition 4 (Total order of threshold contracts)**.** *Fix $\theta \in (0,1)$. Let t be the n-threshold contract and t' be the n'-threshold contract such that both threshold contracts induce θ. Contract t has a strictly lower expected cost to the firm if $n > n'$.*

Figure 1.1 provides an example to illustrate the intuition of the result. In the example, the firm incentivizes a customer with four referral friends to choose an effort level of 0.8. Each curve is the marginal rewards curve of a different threshold contract, with the expected referral reward obtained from the area under the curve from 0 to the target effort level. By comparing the areas under the curve, we see that higher-threshold contracts are more cost-effective in incentivizing this particular target effort level. The reason for the clear graphical comparison is that for every pair of curve, the curve of the highest-threshold contract always single-crosses that of a lower-threshold contract exactly at the target effort level on the open unit interval. It turns out that this single-crossing property holds between every pair of threshold contracts for every interior target effort level. To summarize this single-crossing property, we say that an effort-induced distribution $F(\cdot;\cdot)$ is said to have the **threshold single-crossing property** (TSCP) if $\frac{f_q(n;q)}{F_q(n;q)}$ is non-decreasing in $q \in (0,1)$ for $0 < n < N$. The effort-induced binomial distribution satisfies strict TSCP. Section 1.6 discusses TSCP as a condition for highest-threshold optimality in detail.

TSCP has some nice interpretable economic insights. Marginal rewards can be interpreted as a measure of responsiveness of a contract's referral rewards with respect to extra efforts. With this interpretation, TSCP implies that at low effort levels, the marginal rewards of low-threshold contracts are high relative to high-threshold contracts. In other words, when effort levels are low, the customer is able to gain relatively large extra rewards if working a little harder, whereas

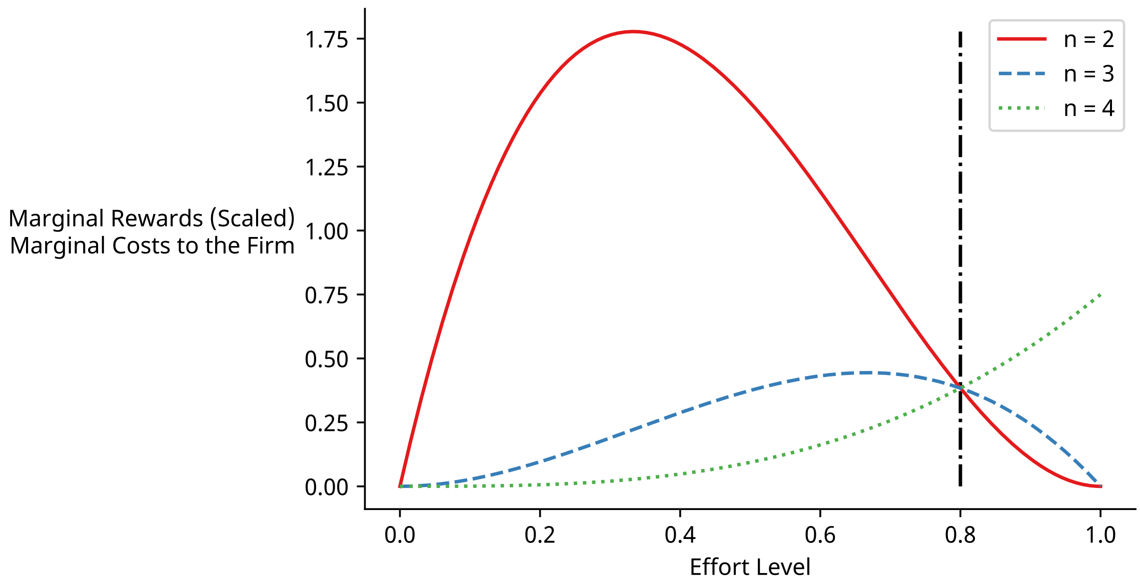

Figure 1.1: Marginal rewards Curves of Threshold Contracts with $N = 4$

the extra rewards are small in higher-threshold contracts. Consequently, in a lower threshold contract, the customer is able to get a relatively large expected referral reward even if the chosen effort level is still far below the target effort level. In contrast, in a higher-threshold contract, the customer's expected referral reward is low unless the chosen effort level is sufficiently close to the target. This difference makes higher-threshold contracts better approximation of effort-dependent contracts than lower-threshold contracts, hence making higher-threshold contracts more cost-effective.

1.3.3 Threshold vs Linear Contracts

Since linear contracts are popular among referral programmes, I compare the threshold contracts with linear contracts based on their cost-effectiveness. One property of a linear contract b is that $v_q(b; q) = \Delta b_1 = \cdots = \Delta b_N$, where $\Delta b_n = b_n - b_{n-1}$ denotes the incremental reward in each outcome. Given the convexity of $c(\cdot)$, for each interior effort level, there exists a linear contract inducing this effort level. It turns out that given a fixed interior effort level, the threshold contracts, except the highest-threshold contracts, are not always more cost-effective than the linear contracts.

Proposition 5 (Cost-Effectiveness comparison between threshold and linear contracts)**.** *Fix $\theta \in (0, 1)$ and assume there exists a contract in which the customer optimally chooses θ. There exist $0 = q_1^* \leq \cdots \leq q_N^* = 1$ such that it is (weakly) more cost-efficient to use an n-threshold contract to induce θ than to use a linear contract if and only if $\theta \leq q_n^*$. If for some n, $q_n^* < 1$, then $q_{n+1}^* > q_n$.*

Here is some intuition for the cost-effectiveness comparison between the linear contracts and the highest-threshold contracts. Figure 1.2 draws the marginal rewards curve of two threshold contracts and the linear contract when $N = 4$ and the target effort level is $\theta = 0.6$. Since the curve

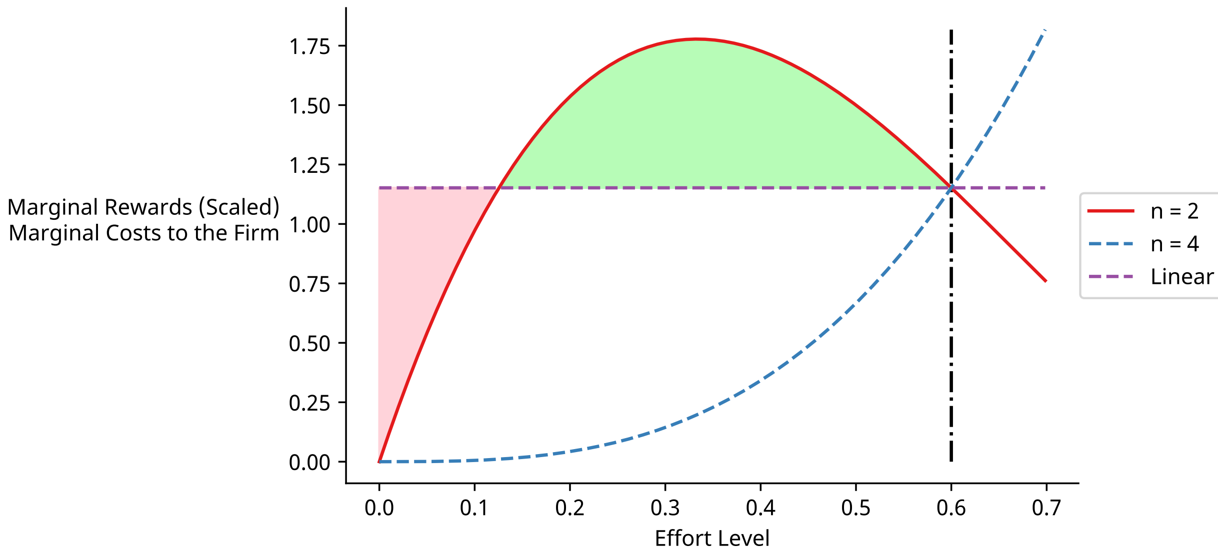

Figure 1.2: Marginal rewards Curves of Threshold and Linear Contracts with $N = 4$

of the highest-threshold contract is increasing, it is more cost-effective than the linear contract. The comparison between the 2-threshold contract and the linear contract, however, is unclear at $\theta = 0.6$. Since the curve of the 2-threshold contract is increasing for small efforts, the 2-threshold contract is more cost-effective than the linear contract when the target effort level is low: in the pink shaded region (on the left), the 2-threshold contract is increasingly more cost-effective than the linear contract. As the effort level gets larger, however, the cost-effective advantage of the 2-threshold contract wanes: when the area of the green shaded region (on the right) is more than the pink region, then it is the linear contract that is more cost-effective than the 2-threshold contract. The proof shows that there exists a unique cut-off such that a threshold contract is more cost-effective than a linear contract if and only if the target effort level does not exceed this cut-off.

1.4 Extensions of Baseline Model

This section analyses a number of extensions that are highly relevant to the design of referral programmes. In these extensions, the highest-threshold contract is not always optimal, but the inclusion of threshold contracts still acts as a powerful device for incentivizing efforts.

1.4.1 Positive Reserve Utility

One anecdotal observation from referral programmes is that it appears costly to have the customer get started on referrals: a customer often does not respond to a referral programme at all. One way to interpret this observation is that the customer's reserve utility is strictly positive but the referral programme does not offer enough payoffs. To model this interpretation, assume the customer's reserve utility is $r \geq 0$. Consequently, the firm's optimization problem in (1.2) is changed to

$$\max_{b\in\mathbb{R}^{N+1}, q\in[0,1]} \quad U(b;q) = \pi N q - v(b;q) \tag{1.3}$$

$$\text{s.t.} \quad v(b;q) - c(q) \geq r, \tag{1.3a}$$

$$v(b;q) - c(q) \geq v(b;q') - c(q'), \quad \forall q' \in [0,1], \tag{1.3b}$$

$$b_n \geq 0, \qquad 0 \leq n \leq N. \tag{1.3c}$$

When the customer's participation constraint binds at optimum, an optimal contract may not necessarily be an N-threshold contract because the class of N-threshold contract may be too cost-effective such that the firm becomes concerned about the customer overworking. For example, let $t \in \mathbb{R}^{N+1}$ be the N-threshold contract such that $v_q(t;q_{FB}) = c'(q_{FB})$.[12] Assume $v(t;q_{FB}) - c(q_{FB}) < r$. If the firm still wants to use an N-threshold contract, it needs to raise t_N to give the customer more surplus, but after q exceeds q_{FB}, doing so would lower the firm's profit because the firm's marginal revenue with respect to efforts, which is πN, would be lower than the marginal rewards the firm needs to pay the customer, which is $v_q(t;q)$. Moreover, if there exists a contract in which the customer chooses q_{FB} and the participation constraint binds, this contract is optimal to the firm since the first-best outcome is implemented.

Kim (1997) deals with the continuous-outcome case and discusses sufficient conditions for implementing first-best outcome. That paper finds the first-best-implementing contract to be a bonus contract: the customer gets a fixed additional reward for each additional new referral and gets a fixed bonus reward for exceeding a threshold. In relation to this paper, I characterize the optimal contract for different reserve utility, and show that when the reserve utility is moderate, there exists a bonus contract that implements the first-best outcome.

Proposition 6 (Positive reserve utility)**.** *Assume there exists a strictly profitable contract with respect to* (1.3)*. There exists $\underline{r} < \bar{r}$ such that*

- *if $r \leq \underline{r}$, then every optimal contract is an N-threshold contract.*
- *if $r \in (\underline{r}, \bar{r})$, then there exists an optimal contract as a strict combination of a linear and an N-threshold contract. Moreover, the first-best outcome is implemented.*
- *if $r \geq \bar{r}$, then it is optimal for the firm not to have a referral programme.*

Since the firm is not allowed to charge the customer in the referral programme, the customer gets a strictly positive payoff in an N-threshold contract. As a result, when the reserve utility is low, it is business as usual to the firm as the optimal N-threshold contract provides sufficient payoff to the customer. When the reserve utility is excessively high, then it makes no sense for the firm to have a referral programme even in the first-best case.

When the reserve utility is moderate, given the effort level at optimum, if the firm attempts to use a linear contract, then the contract is able to satisfy both the effort-choice and participation

[12] It is assumed here that the first-best effort level stays the same, i.e., r is assumed not to be too large that the first-best effort level is zero.

constraint, but it turns out that the linear contract alone is not cost-effective. On the other hand, if the firm only uses an N-threshold contract, then the customer's effort-choice constraint holds, but the participation constraint does not. What can be done here is to mix the linear and the N-threshold contract together and the proof shows that there exists a unique combination that cost-efficiently incentivizes the target effort level.

Note that this result about the first-best implementation is an existence result. For every pair of contracts such that one contract induces q_{FB} in (1.3) and the other one induces q_{FB} in (1.2) but violates (1.3a), then by the similar reasoning in the proof of Proposition 6, some convex combination of this pair of contracts implements the first-best outcome.

1.4.2 Limited Liability of Firm

In a highest-threshold contract, the firm pays a large referral reward with a small probability. Sometimes this highest-threshold reward can be larger than the total referral revenue, as the following example shows.

Example 1. Consider the case with $N = 2$, $\pi = 1$, and $c(q) = 6q^3$. The optimal N-threshold contract gives the customer a reward of $\sqrt{6} > 2\pi$ when both friends are referred. Hence the firm's limited liability condition would be violated in this contract.

In the optimal N-threshold, the firm gets the entire surplus for most of the times. To incentivize the customer to exert the desired effort level, it is possible that the firm is willing to lose some money in some outcomes with very small probabilities. However, in practice, it is common for the firm to set a budget constraint on its referral programme. One common example is that the firm requires its referral programme to never run into a fiscal deficit in every referral outcome, i.e., $b_n \leq \pi n$ for all n. In the example above, the optimal N-threshold contract is no longer feasible as the budget constraint is violated for the highest referral outcome. To reflect this particular budget constraint requirement, the firm's profit-maximization problem is changed to

$$\max_{\substack{b\in\mathbb{R}^{N+1} \\ q\in[0,1]}} \quad \pi N q - v(b;q) \tag{1.4}$$

$$\text{s.t.} \quad v(b;q) \geq c(q) \tag{1.4a}$$

$$v(b;q) - c(q) \geq v(b;q') - c(q') \quad \forall q \in [0,1] \tag{1.4b}$$

$$0 \leq b_n \leq \pi n \quad \forall 0 \leq n \leq N \tag{1.4c}$$

This newly imposed assumption, together with the limited liability of customer, leads to a result with a similar flavour to the main result in Innes (1990). By Proposition 4, higher threshold contracts are more cost-effective than lower threshold contracts. One may conjecture that the upper bounds of referral rewards for higher outcomes should bind at the optimum. The following result shows that this is indeed the case.

Theorem 1 (Optimal contract with limited liability of firm)**.** *Assume the firm has limited liability, i.e.,* (1.4c) *holds. An optimal contract exists, i.e.,* (1.4) *has a solution. If a contract that is strictly profitable with respect to* (1.4) *exists, then every optimal contract b has the property that $b_{n_1} > 0$ implies $b_{n_2} = \pi n_2$ for every $n_2 > n_1$.*

In Innes (1990), since it is the agent that writes the contract, the agent would like to use the most costly way to incentivize himself, leading to payments concentrated in the lowest outcomes, whereas in this paper, as it is the firm that writes the contract, the goal of the contract instead becomes using the least costly way to incentive efforts, leading to payments only in the highest outcomes. Also, in Innes (1990), every positive payment binds the limited-liability constraint because the agent in that model has an incentive to put a unit mass at the lowest outcome if no upper bounds on payment exist.

The limited liability of firm assumption has an extended interpretation that the firm's division running the referral programme has an overall budget $\overline{\pi} \geq 0$ such that for every contract b, $b_n - \pi n \leq \overline{\pi}$. The same proof in Theorem 1 can be adapted to show that in every optimal contract, if $b_n > 0$, then $b_{n+1} - \pi n = \overline{\pi}$. The case with limited liability of firm corresponds to the case with $\overline{\pi} = 0$; the base model without any limited liability of firm can be represented by $\overline{\pi} = \infty$ or $\overline{\pi}$ being very large. The following result states that relaxing the budget constraint eventually leads to every optimal contract being an N-threshold contract.

Proposition 7 (Optimal contract with more flexible budget constraints for the firm)**.** *Fix $\theta > 0$ and assume there exists a contract in which the customer chooses θ. Let $\overline{\pi}$ be the budget constraint of the firm such that for every contract b, $b_n - \pi n \leq \overline{\pi}$. Let $b(\overline{\pi})$ be the cost-efficient contract inducing π and let n^* be the smallest outcome such that $b_{n^*}(\overline{\pi}) > 0$. For another budget constraint $\overline{\pi}' > \overline{\pi}$, $n^*(\overline{\pi}') \geq \overline{\pi}$; for $\overline{\pi}'$ large enough, $n^*(\overline{\pi}') = N$.*

Proof. From the proof of Theorem 1, for every $n \geq n^*(\overline{\pi})$, $f_q(n;\theta) > 0$. Therefore, by Theorem 1, for $\overline{\pi}' > \overline{\pi}$, if $b_{n^*(\overline{\pi})}(\overline{\pi}') \geq b_{n^*(\overline{\pi})}(\overline{\pi})$, then $v_q(b(\overline{\pi}');\theta) > c'(\theta)$, a contradiction. Therefore, $n^*(\overline{\pi}') \geq n^*(\overline{\pi})$. Lastly, for $\overline{\pi}'$ large enough, the N-threshold contract t inducing θ has $t_n - \pi n \leq \overline{\pi}'$, and thus $n^*(\overline{\pi}') = N$ for $\overline{\pi}'$ large enough. □

1.4.3 Risk-Averse Customer

The risk-neutrality of the customer contributes to the baseline model's tractability. This section digress to the analysis with a risk-averse customer. Specifically, assume the customer's aggregate utility is $V(x) - c(q)$, where x is a referral reward and $V(\cdot)$ is the customer's utility of rewards. Assume $V(0) = 0$, $V'(x) > 0$ and $V''(x) \leq 0$ for all x. The firm's profit-maximization problem is

$$\max_{\substack{b \in \mathbb{R}^{N+1} \\ q \in [0,1]}} \quad \pi N q - \sum_{n=0}^{N} b_n f(n;q) \tag{1.5}$$

$$\text{s.t.} \quad \sum_{n=0}^{N} V(b_n) f(n;q) - c(q) \geq 0 \tag{1.5a}$$

$$\sum_{n=0}^{N} V(b_n) f(n;q) - c(q) \geq \sum_{n=0}^{N} V(b_n) f(n;q') - c(q'), \quad \forall q' \in [0,1] \tag{1.5b}$$

$$b_n \geq 0, \quad 0 \leq n \leq N \tag{1.5c}$$

The existence of an optimal contract can be established by a similar reasoning to that in Theorem 2 proved in Section 1.6, which is omitted to avoid repetition. Additionally, by treating the customer's utility from rewards as referral rewards, it can be shown that for each $1 \leq n \leq N$ and an interior effort level, there exists an n-threshold contract inducing this effort level.

It turns out that in when the customer is risk-averse, it is still useful for the firm to set a threshold for the customer by not paying the customer anything if the referral outcome is below the threshold.

Proposition 8 (No rewards in dis-incentivizing outcomes)**.** *Let b be an optimal contract and assume the customer chooses $\theta \in (0,1)$ in this contract. For each $0 \leq n \leq N$, if $\theta > \frac{n}{N}$, then $b_n = 0$.*

Sketch of Proof. For each n, $f_q(n;q) < 0$ if and only if $q > \frac{n}{N}$, i.e., for $q > \frac{n}{N}$, more efforts would make outcome n less likely. The proof shows that if in an effort level θ chosen by the customer in an optimal contract $f_q(n;\theta) < 0$, then $b_n = 0$. □

The result has an intuitive interpretation that the firm pays the customer to exert more efforts instead of dis-incentivizing efforts, which is the case when $\theta > \frac{n}{N}$. The following result uses the sufficient condition derived in Grossman and Hart (1983) to show that optimal contracts are convex.

Proposition 9 (Convex optimal contract)**.** *Assume $1/V'(\cdot)$ is concave. Let b be an optimal contract and the customer exerts effort level $\theta \in (0,1)$, if $b_n > 0$, then $b_{n+2} - b_{n+1} \geq b_{n+1} - b_n \geq 0$, i.e., every optimal contract is non-decreasing and convex.*

The proof shows that at optimum, the customer marginal monetary equivalent with respect to utility is non-decreasing in the number of referrals. Since $1/V'(\cdot)$ is assumed to be positive and concave, this implies that the incremental rewards must be non-decreasing for each outcome and hence every optimal contract is non-decreasing and convex.

1.4.4 Communication Cost

The baseline model assumes that the customer's effort cost only depends on the conversion rate and is invariant to the number of referral messages. In practice, sending the referral message itself can be costly. One immediate interpretation of this per-message communication cost is the inconvenience that comes with sending a referral message. A more extended interpretation, which makes this subsection more powerful, is the latent social cost: if the customer recommends a product to a friend but the friend does not like the product, the customer may suffer some social cost for recommending a "bad" product.

To analyse how the firm's referral programme design problem changes with this new communication cost, assume the customer incurs a fixed communicate cost $s \geq 0$ whenever the customer tries to refer a friend. The extended cost structure in this subsection covers that in Lobel, Sadler, and Varshney (2017), which only includes communication cost. If efforts and the number of referral messages were contractible, the firm's optimization problem would be

$$\max_{\substack{\beta\in\mathbb{R},q\in[0,1]\\ 0\leq\tilde{N}\leq N}} \pi q\tilde{N} - \beta, \quad \text{s.t.} \quad \beta - c(q) \geq s\tilde{N}. \tag{1.6}$$

Assume (1.6) admits a solution such that $q > 0$ and the objective function is strictly positive. Since marginal revenue with respect to efforts is increasing in $\tilde{N}$, if $q > 0$ at optimum, then $\tilde{N} = N$ in every solution to (1.6).

When neither efforts nor the number of referral messages are contractible, the firm's optimization problem is

$$\max_{\substack{b\in\mathbb{R}^{N+1},\\ q\in[0,1],\\ 0\le\tilde{N}\le N}} \quad U(b;q,\tilde{N}) = \pi\tilde{N}q - v(b;q,\tilde{N}) \tag{1.7}$$

$$\text{s.t.} \quad v(b;q,\tilde{N}) - c(q) - s\tilde{N} \geq 0, \tag{1.7a}$$

$$v(b;q,\tilde{N}) - c(q) - s\tilde{N} \geq v(b;q',N') - c(q') - sN', \quad \forall q' \in [0,1], 0 \leq N' \leq N \tag{1.7b}$$

$$b_n \geq 0, \quad 0 \leq n \leq N. \tag{1.7c}$$

Although this paper does not provide a characterization of the solution to (1.7), it can be shown that using or including a threshold contract strictly improves the cost-effectiveness over using only a linear contract.

Proposition 10. *For every optimal linear contract, there exists a combination of linear contract and an N-threshold contract such that the firm's profit is higher in the combined contract.*

The proof first shows that in an optimal linear contract, it is necessary that the customer willingly sends N referral messages. Given this result, by treating the communication cost of sending N referral messages as the customer's reserve utility, by Proposition 6, the optimal contract must be either a N-threshold contract or a combination of a linear and N-threshold contract, which is a strict improvement over the linear contract if the alternative contract indeed incentivizes N messages.

Assume the optimal contract is an N-threshold contract, since the customer gets nothing for sure if sending $\tilde{N} < N$ messages, the customer has an incentive to send N messages. Now assume the optimal contract is a combination of a linear and an N-threshold contract. If the customer sends $\tilde{N} < N$ messages, then contract is effectively a linear contract. By the earlier discussion on linear contract, the customer has an incentive to send N messages without the N-threshold contract. Since the N-threshold contract only gives the customer more incentive to send N messages, the customer has an incentive to send N messages in the combined contract.

1.5 Heterogeneity in Number of Referrable Friends

In practice, the firm may only have incomplete information about the number of referral friends. To account for this motivation, this section assumes that a customer's number of referrable friends is private to the customer. For simplicity, assume that there are two types of customers: the low type has $N_1 > 0$ friends, whereas the high type has $N_2 > N_1$ friends. Let $\alpha \in (0,1)$ denote the proportion of customers that are low type and $1-\alpha$ the proportion of the high type.

Given a contract $b \in \mathbb{R}^{N+1}$, let $V^*(b;N) = \max_{q\in[0,1]} v(b;q,N') - c(q)$, the maximal utility the agent can get if the customer takes up the contract. Given the continuity of $v(b;q,N)$ and $c(q)$ in both b and q, $V^*(b; N)$ is continuous in b. Given a contract b, define

$$q^*(b; N) = \max\left\{\arg\max_{q\in S} \pi qN - v(b; q, N)\right\}, \quad \text{where } S = \arg\max_{q'\in[0,1]} v(b; q', N), \tag{1.8}$$

which is the maximal effort level that maximizes the firm's profit among effort levels that are optimal to a customer with N friends in b.

1.5.1 With screening

First assume that the firm can offer a **menu of referral contracts** to customers. I focus on direct mechanisms in which a customer reports the customer's type, which is the number of referrable friends, to get a corresponding referral contract. This setup is a model with both hidden actions and hidden information.[13]

A menu $(b^1, b^2) \in \mathbb{R}^{N_2+1} \times \mathbb{R}^{N_2+1}$ is said to be **incentive compatible** if $V^*(b^1; N_1) \geq V^*(b^2; N_1)$ and $V^*(b^2; N_2) \geq V^*(b^1; N_2)$, i.e., each type has no incentive to choose the contract for a different type.[14] Given an incentive-compatible menu $(b^1, b^2) \in \mathbb{R}^{N_2+1} \times \mathbb{R}^{N_2+1}$, define

$$\begin{aligned}
\Pi(b^1, b^2) = &\max_{\substack{q_1\in[0,1],\\ q_2\in[0,1]}} \alpha U(b^1; q_1, N_1) + (1-\alpha)U(b^2; q_2, N_2) \\
&\text{s.t.} \quad v(b^1; q_1, N_1) = V^*(b^1; N_1) \\
&\qquad\;\; v(b^2; q_2, N_2) = V^*(b^2; N_2),
\end{aligned}$$

which is the maximal profit the firm can obtain when both types choose effort levels optimally. Under this setup, the firm's profit-maximization problem is

$$\max_{\substack{b^1\in\mathbb{R}^{N_2+1}, b^2\in\mathbb{R}^{N_2+1}\\ q_1\in[0,1], q_2\in[0,1]}} \alpha U(b^1; q_1, N_1) + (1-\alpha)U(b^2; q_2, N_2) \tag{1.9}$$

$$\text{s.t.} \quad v(b^i; q_i, N_i) \geq c(q_i) \qquad 1 \leq i \leq 2 \tag{1.9a}$$

$$v(b^i; q_i, N_i) - c(q_i) \geq v(b^j; q', N_i) - c(q') \qquad 1 \leq i \leq 2, 1 \leq j \leq 2, \tag{1.9b}$$

$$b^i_n \geq 0 \qquad 1 \leq i \leq 2, 0 \leq n \leq N_i. \tag{1.9c}$$

A menu (b^1, b^2) is said to be **strictly profitable** with respect to (1.9) if there exists a pair of effort levels (q_1, q_2) such that the tuple (b^1, b^2, q_1, q_2) is feasible in (1.9) and $\alpha U(b^1; q_1, N_1) + (1 - \alpha)U(b^2; q_2, N_2) > 0$. A menu is said to be **optimal** if there exists (q_1, q_2) such that (b^1, b^2, q_1, q_2) solves (1.9).

Let $t^l \in \mathbb{R}^{N_2+1}$ denote an N_1-threshold contract that is optimal when there is only the low type and $t^h \in \mathbb{R}^{N_2+1}$ an N_2-threshold contract that is optimal when there is only the high type. Since

[13] A weakly more general modelling is to have a customer report both the type and effort level. These two ways of modelling are outcome-equivalent in equilibrium since in equilibrium, customers of the same type choose the same contract, which would translate into a contract for that type in the case where the customers only report their types.

[14] Let $b^1 \in \mathbb{R}^{N_2+1}$ so that the utility of the high type choosing b^1 is defined.

$N_2 > N_1$, the low type has no incentive to choose t^h. For the high type, if $V^*(t^h; N_2) \geq V^*(t^l; N_2)$, then (t^l, t^h) is incentive compatible. Since t^h and t^l are the optimal contract to the firm when there is only type, (t^l, t^h) is optimal if it is incentive compatible. However, the menu above may not be incentive compatible, in which case the analysis is not straightforward. Partial characterizations of optimal menus are provided. The following result establishes the existence of an optimal menu.

Proposition 11 (Positive efforts at optimum)**.** *An optimal menu exists, i.e.* (1.9) *has a solution. If there exists a menu that is strictly profitable with respect to* (1.9)*, then in every optimal menu, both types choose strictly positive effort levels.*

The rest of this section assumes the existence of a contract that is strictly profitable with respect to (1.9). That both types exert strictly positive efforts implies that there is no exclusion of the low type in an optimal menu whenever $\alpha > 0$, which is partially due to the limited liability of the customer. Indeed, if only the high type exerts a positive effort level in an optimal menu, then from the single-type case, the firm would necessarily offer an N_2-threshold contract to the high type, from which the high type would get a strictly positive payoff. The proof shows that the firm can further improve the profit by giving the low type a contract that is near $0 \in \mathbb{R}^{N_2+1}$.

The following result provides a condition under which the contract for the high type must be an N_2-threshold contract in every optimal menu.

Proposition 12 (Threshold optimality for high type)**.** *Let $t^l \in \mathbb{R}^{N_2+1}$ be an optimal contract when there is only the low type. If there exists some N_2-threshold contract $t \in \mathbb{R}^{N_2+1}$ such that $V^*(t; N_2) = V^*(t^l; N_2)$ and $q^*(t; N_2) \leq q_{FB}(N_2)$, then in every optimal contract (b^1, b^2), b^2 is an N_2-threshold contract with $q^*(b^2; N_2) \leq q_{FB}(N_2)$, i.e., b^2 induces the high type to choose an effort not exceeding the first-best effort level.*

The proof shows directly how the firm can strictly improve the profit if the contract for the high type is not an N_2-threshold contract. Specifically, the firm can provide an N_2-threshold contract such that either the high type exerts the same effort level yet with strictly lower expected rewards or the high type becomes indifferent between the two contracts in the new menu and chooses a strictly higher effort level, which brings more revenues to the firm as long as the effort level is less than the first-best level. Indeed, the condition that $q^*(t; N_2) \leq q_{FB}(N_2)$ is crucial. To see this, consider the case where α is close to 1. In this case, intuitively, the firm would like the contract for the low type to be close to t^l. If $q^*(t; N_2) \leq q_{FB}(N_2)$ in Proposition 12, then in contract t, both firm and the customer's payoffs are increasing in $q \leq q^*(t; N_2)$. However, when the condition does not hold, then the firm's profits are decreasing in $q > q_{FB}(N_2)$, i.e., the firm would like the customer to choose a lower effort level.

The result that the contract for the high type is necessarily an N_2-threshold contract contrasts with other models with both hidden actions and hidden information. For example, Laffont and Tirole (1986), McAfee and McMillan (1987), Rao (1990), and Melumad and Reichelstein (1989) consider cases with a continuum of types and outcomes. In all of these papers, menus of linear contracts are shown to be optimal, whereas, in this paper, this class of menus is strictly suboptimal. Here is a short intuition for the difference in the results. In the models mentioned above, if the contract for the high type is replaced with a strictly more cost-effective contract such that the high type is induced to either exert the same effort level with strictly lower expected rewards or more efforts that improve the firm's profits, the incentive-compatibility constraint of the low type would not hold. However, in this model, the incentive-compatibility constraint of the low type always holds when the high type's contract is an N_2-threshold contract.

1.5.2 No screening

Given the likely operational cost of maintaining complex referral programmes, which is exogenous to this paper, the firm may wish to have only one referral programme even though there is more than one type. An overall impression from the screening literature is that when there is no screening, then the firm faces a simple choice between the optimal contract when there is only the low type and the one when there is only high type. In this paper, however, the firm's strategic consideration is not straightforward. Indeed, given a contract $b \in \mathbb{R}^{N_2+1}$, any changes to b_n for $n > N_1$ is not going to affect the low type's incentives with respect to the contract. Therefore, the firm needs to take into account the incentives of both types in the referral contract design.

This subsection imposes an additional assumption that the firm commits to providing only non-decreasing referral contracts, i.e., $b_0 \leq b_1 \leq \cdots \leq b_N$. The assumption is reasonable for referral programmes since there are few actual referral programmes that are non-decreasing. One possible conjecture in case that the firm does take away rewards from the customer for more referrals is that a customer would ask the friends whom the customer has successfully converted to let the customer know before going to the firm so that the customer could choose a subset of friends to report to the firm as the customer's referrals. Under this conjecture, the firm would then have no incentives to have a contract in which rewards could decrease when the number of referrals rises. Let $\tilde{U}(b; q_1, q_2) = \alpha U(b; q_1, N_1) + (1-\alpha) U(b; q_2, N_2)$. The firm's profit-maximization problem is therefore

$$\max_{\substack{b \in \mathbb{R}^{N_2+1} \\ q_1 \in [0,1], q_2 \in [0,1]}} \quad \tilde{U}(b; q_1, q_2) = \alpha U(b; q_1, N_1) + (1-\alpha) U(b; q_2, N_2) \tag{1.10}$$

$$\text{s.t.} \quad v(b; q_i, N_i) \geq c(q_i) \quad 1 \leq i \leq 2 \tag{1.10a}$$

$$v(b; q_i, N_i) \geq v(b; q', N_i) \quad \forall q' \in [0,1], \quad 1 \leq 1 \leq 2 \tag{1.10b}$$

$$0 \leq b_0 \leq b_1 \leq \cdots \leq b_{N_2} \tag{1.10c}$$

In this subsection, a contract b is said to be **optimal** if there exists (q_1, q_2) such that (b, q_1, q_2) solves (1.10). Given a contract b, $\Pi(b, b) = \tilde{U}(b; q^*(b; N_1), q^*(b; N_2))$ is the maximal profit the firm can get from b if the two types choose their effort levels optimally. The following result establishes the existence of an optimal contract.

Proposition 13 (Positive efforts for high type at optimum)**.** *An optimal contract exists, i.e.* (1.10) *has a solution. If there exists a contract that is strictly profitable with respect to* (1.10)*, then in every optimal contract, the high type always chooses a strictly positive effort level.*

For $1 \leq n_1 < n_2 \leq N_2$, a contract $b \in \mathbb{R}^{N_2+1}$ is said to be an (n_1, n_2)-threshold contract if $b_0 = b_1 = \cdots = b_{n-1} < b_{n_1} = b_{n_1+1} = \cdots = b_{n_2-1} \leq b_{n_2} = b_{n_2+1} = \cdots = b_{n_2}$. Since setting $b_{n_1} = b_{n_2}$ gives a contract with only one threshold, the set of contracts with a single threshold is included in the set of contracts with two thresholds. In the following result, the optimal contract with two types is shown to be a contract with two thresholds.

Lemma 1 (Cost-Efficiency of contracts with two thresholds). *Fix* $(\theta_1, \theta_2) \in [0,1]^2 \setminus (0,0)$*. Assume there exists a contract* b *such that* (b, θ_1, θ_2) *is feasible in* (1.10)*. There exists a contract with two thresholds that is cost-efficient and induces* (θ_1, θ_2)*.*

Sketch of Proof. Consider the relaxed problem of the cost-minimization part of (1.10), i.e., the effort-choice constraints (1.10b) are replaced by the relaxed first-order conditions

$$v_q(b; \theta_i, N_i) = c'(\theta_i), \quad 1 \le i \le 2.$$

The relaxed problem has the same set of solutions with the actual problem (1.10) because by Proposition 3, the first-order condition uniquely characterizes the customer's effort-choice constraints in non-decreasing contracts. The constraint substitutions make the firm's cost-minimization problem a linear programming. With results from linear programming, there exists a contract with two thresholds that is a solution to the relaxed cost minimization problem (1.17). Since the two types' effort choice constraints hold, the solution to the relaxed problem (1.17) also solves the actual cost minimization part of (1.10). □

The following result provides a partial characterization of an optimal contract. Specifically, it can be shown that whenever the higher threshold is above n_1, then setting the second threshold to N_2 improves the firm's profits.

Proposition 14. *Assume the existence of a contract that is strictly profitable with respect to* (1.10)*. There is an* (n_1, n_2)*-threshold contract for some* $1 \le n_1 < n_2 \le N_2$ *such that it is optimal. If* $n_2 > N_1$ *and* $b_{n_2} > b_{n_1}$*, then* $n_2 = N_2$*.*

Sketch of Proof. For every contract b that induces (θ_1, θ_2), since the contract is assumed to be non-decreasing, $b_n > 0$ for $n > N_1$ always induces the high type to exert more efforts than when b_n is set to 0 for all $n > N_1$. The proof shows that to induce the high type to exert the extra efforts due to payments b_n for $n > N_1$, setting $b_n = 0$ for $N_1 < n < N_2$ and $b_{N_2} > 0$ is the cost-efficient way to do so. □

The search over contracts with two thresholds such that both thresholds are below N_1 has the computational costs $O(N_1^2)$; the search over contracts with two thresholds with the higher threshold being N_2 has computational costs of $O(N_1)$. Therefore, given the result on the form of optimal contracts, the firm's search for an optimal contract has cost $O(N_1^2)$.

Remark 2. If a strictly profitable menu exists, Proposition 11 implies that menus of a single N_2-threshold contract only for the high type are not optimal. Furthermore, if the condition of Proposition 12 holds, then every optimal menu necessarily has two distinct non-zero contracts. Therefore, the firm is strictly better off with screening than without screening.[15]

[15] Melumad and Reichelstein (1989) show that whether screening is strictly better depends on how costly it is for the principal to screen different types. In this model, the condition of Proposition 12 can be considered as a requirement that it is not too costly to use an N_2-threshold contract to screen: if the N_2-threshold used to incentivize the high type away from t^l in Proposition 12 is not too costly, then screening is valuable to the firm.

1.6 Discussions

This section discusses the sufficient conditions for the highest-threshold optimality result in the baseline model. The firm's profit-optimization problem is still characterized by (1.2),[16] but the binomial distribution is not imposed.

Let $F(\cdot; q, N)$ be the cumulative distribution function of referral number with $f(\cdot; q, N)$ being its probability mass function when the existing with N referrable friends chooses effort level $q \in [0, 1]$. There are three distributional assumptions we discuss as (part of) sufficient conditions for the highest-threshold optimality result, formally defined below.

Definition 2. The effort-induced distribution $F(\cdot; \cdot)$ is said to have

(a) the threshold single-crossing property (TSCP) if $\frac{f_q(n;q)}{F_q(n;q)}$ is non-decreasing in $q \in (0, 1)$ for $0 < n < N$.

(b) the monotone likelihood ratio property (MLRP) if $\frac{f_q(n;q)}{f(n;q)}$ is non-decreasing in n for $0 \le n \le N$.

(c) the decreasing hazard rate property (DHRP) if $\frac{f_q(n;q)}{1-F(n;q)}$ is non-increasing in $q \in (0, 1)$ for $0 \le n < N$.

Balmaceda et al. (2016) show the optimality of highest-threshold contracts with MLRP. The proof technique in Poblete and Spulber (2012), who deal with continuous outcomes, can be adapted to this paper's setup to show the same result with DHRP, which MLRP implies. However, Poblete and Spulber (2012) make the additional assumption that the referral contract must be non-decreasing. In contrast, this paper shows that under the discrete-outcome setting, the monotonicity assumption is not necessary.

To obtain the threshold-optimality result as in Proposition 2, we first make the following regularity conditions about the distribution.

Assumption A1 (Distributional regularity)**.**

(a) $F(n; q)$ is continuously differentiable in $q \in (0, 1]$ for $0 \le n \le N$.

(b) $f(n; q) > 0$ for $q \in (0, 1)$ and $0 \le n \le N$.

(c) $F(0; 0) = 1$.

(d) $F_n(n; q) < 0$ for $0 \le n < N$ and $q \in (0, 1)$.

Among these assumptions, A1(b) is made mainly to ensure that DHRP and MLRP are well-defined; A1(c) ensures that for the firm to have a strictly positive effort, the customer must exert strictly positive efforts. Exerting efforts q incurs a cost of $c(q)$ to the customer. The following regularity assumption is imposed on the effort cost function. (A1(d)) is the condition that higher

[16] Note that the constraints of (1.2) implicitly impose the limited liability of customer assumption.

efforts first-order stochastically dominate lower efforts. Note that this assumption is implied by both MLRP and DHRP, and it is only necessary for the highest-threshold-optimality result for DHRP.

Assumption A2 (Cost regularity)**.** $c(0) = 0$, $c'(q) > 0$ and $c''(q) > 0$ for $q \in (0, 1)$.

The following assumption ensures that in every solution to (1.2), the contract is non-zero.

Assumption A3 (Existence of strictly profitable contract)**.** There exists (b, q) that is feasible in (1.2) and $U(b, q) > 0$.

Before characterizing optimal contracts with respect to (1.2), it is necessary to clarify that (1.2) has a solution.

Theorem 2 (Existence of optimal contract)**.** *Assume A1 and A2 hold. An optimal contract exists, i.e.,* (1.2) *has a solution. If further A3 holds, then in every solution* (b, q) *to* (1.2)*,* $q > 0$*.*

The assumptions so far, together with DHRP and IMCP, are sufficient for an optimal contract to be a highest-threshold contract. In contrast to Poblete and Spulber (2012), the result does not need the assumption that the firm must provide a non-decreasing contract.

Theorem 3 (Highest-Threshold optimality with DHRP)**.** *Assume IMCP and A1-A3 hold, and* $F(\cdot;\cdot)$ *has strict DHRP. There exists a solution* (b, q) *to* (1.2) *in which* b *is an* N*-threshold contract. Moreover, if* $q < 1$*, then* b *must be an* N*-threshold contract.*

The proof has two steps. Firstly, it shows that for every interior target effort level, it is uniquely efficient to use a highest-threshold contract to incentivize that effort level. The proof then proceeds to use linear programming to show that if the cost-efficient contract is not an N-threshold contract, then it can be written as an affine combination of two threshold contracts matching the customer's first-order condition, with the affine weight on the lower-threshold contract being positive. It is shown in the proof that such a combination is not as cost-effective as the highest-threshold contract, leading to the highest-threshold optimality result.

To show MLRP as a sufficient condition for highest-threshold optimality, note that MLRP is known to imply DHRP and hence the highest-threshold optimality result with MLRP is immediately obtained. As for TSCP, it turns out that when A1 holds, then TSCP implies DHRP.

Proposition 15 (TSCP implies DHRP)**.** *If A1 holds, then TSCP implies DHRP.*

With the result above, we can show how TSCP can lead to the same highest-threshold optimality result.

Theorem 4 (Highest-Threshold optimality with TSCP)**.** *Assume IMCP and A1-A3 hold, and* $F(\cdot;\cdot)$ *has strict TSCP. There exists a solution* (b, q) *to* (1.2) *in which* b *is an* N*-threshold contract. Moreover, if* $q < 1$*, then* b *must be an* N*-threshold contract.*

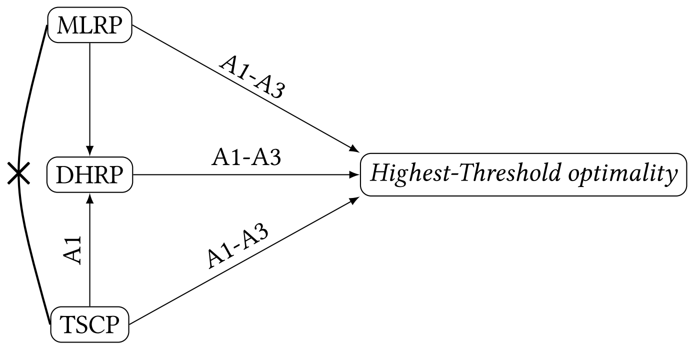

Figure 1.3: Summary of Section 1.6

The result is an immediate consequence of Proposition 15 and Theorem 3, and hence its proof is omitted. The three conditions for highest-threshold optimality, DHRP, MLRP, and TSCP, provide different intuitions, each of each is described below.

MLRP is about the substitution among referral rewards in different referral outcomes. Provided that $f_q(n;q) > 0$ for some $n < N$, MLRP implies $\frac{f_q(n;q)}{f_q(N;q)} < \frac{f(n;q)}{f(N;q)}$. Therefore, the firm can substitute rewards in outcome N for rewards in outcome n so that the customer exerts the same amount of effort yet gets strictly lower expected rewards, and hence the cost-efficiency of highest-threshold contracts.

DHRP has a similar intuition to MLRP, yet the intuition of DHRP is about the substitution among incremental referral rewards in different outcomes. Specifically, DHRP ensures that it is cost-saving for the firm to substitute incremental rewards in outcome N for incremental rewards in every outcome $n < N$.

Lastly, the new TSCP introduced in this paper concerns the change in the responsiveness of referral rewards to the customer's efforts. Specifically, as described in Section 1.3.3, TSCP ensures that in higher-threshold contracts, rewards are relatively unresponsive to extra efforts when the current effort level is low. Consequently, in higher threshold contracts, the customer does not expect to accumulate large rewards until the effort level is high. This observation can be interpreted as a second-best attempt to mimic the first-best contract contingent on effort levels. In the first-best contract where efforts are contractible, the customer does not get paid until the customer's efforts reach the first-best effort level. In the second-best case, the customer is unlikely to get paid anything in a higher threshold contract when the effort level is low. Since in higher threshold contracts the customer gets most of the payoff for the higher part of the customer's efforts, higher-threshold contracts are more cost-effective than lower-threshold contracts as an effort-incentivizing device.

The rest of this section discusses the relations between TSCP, DHRP, and MLRP. In general, DHRP is a strictly weaker than both MLRP and TSCP; TSCP and MLRP do not imply each other. Figure 1.3 summarizes the results in this section.

Between DHRP and TSCP, Proposition 15 has shown that when A1 holds, TSCP implies DHRP. However, in general, TSCP and DHRP do not imply each other. The following example shows that TSCP does not imply DHRP.

Example 2. Fix $N = 2$. Let $f(1;q) = \exp\{\frac{1}{9}q + \frac{1}{9}q^2 - 10\}$ and $f(2;q) = (10-q)\exp\{\frac{1}{9}q + \frac{1}{9}q^2 - 10\}$. This distribution has TSCP but not DHRP (hence not MLRP, either).

Proof. Since $\frac{f(1;q)}{1-F(1;q)} = \frac{1}{10-q}$ is strictly increasing in $q \in (0,1)$. The distribution does not have DHRP. Additionally,

$$\frac{f_q(2;q)}{f_q(1;q)} = 10 - q + \frac{9}{1+2q},$$

which is strictly increasing in $q \in (0,1)$. Hence the distribution has TSCP. □

Even when A1 holds, DHRP is still a strictly weaker condition than TSCP. The following example provides a distribution that has MLRP and hence DHRP, but not TSCP. The example hence also shows that MLRP does not imply TSCP.

Example 3. Fix $N = 2$. Let $f(1;q) = \frac{q}{4}$, $f(2;q) = \frac{5q^2-2q^3}{4}$. Under this distribution, A1 holds. Additionally, this distribution has MLRP and hence DHRP, but not TSCP.

Proof. It is clear that A1(a) holds. Since both $f(1;q)$ and $f(2;q)$ are increasing, A1(d) holds. In addition, since $f(0;0) = 1$ and $f(n;q) \neq 0$ for every $0 \leq n \leq 2$ and $q \in (0,1)$, A1(b) and A1(c) hold. To see that MLRP holds, note that

$$\frac{f(1;q)}{f(2;q)} = \frac{q}{5q^2 - 2q^3} = \frac{1}{5q - 2q^2},$$

which is decreasing in $q \in [0,1]$. Lastly, note that

$$\frac{f_q(1;q)}{f_q(2;q)} = \frac{1}{10q - 6q^2},$$

which is increasing for $q > \frac{5}{6}$, and hence TSCP does not hold. □

Lastly, in the following example, A1 and TSCP hold, which together imply DHRP, but MLRP does not hold. Since MLRP is known to imply DHRP, the following example also shows that DHRP is strictly weaker than MLRP.

Example 4. Fix $N = 3$. Let $f(1;q) = \frac{q}{4}$, $f(2;q) = \frac{2q-q^2}{4}$, $f(3;q) = \frac{q^2}{2}$. The distribution has TSCP and DHRP but not MLRP.

Proof. For MLRP, note that

$$\frac{f(1;q)}{f(2;q)} = \frac{q}{2q - q^2} = \frac{1}{2-q},$$

which is increasing in q and thus $F(\cdot;\cdot)$ does not have MLRP. Since A1 holds, by Proposition 15, DHRP holds if TSCP and A1 hold. Thus it remains to show that TSCP holds. For $f(2;q)$ and $1 - F(2;q)$,

$$\frac{f_q(2;q)}{f_q(3;q)} = \frac{2-2q}{4q} = \frac{1}{2q} - \frac{1}{2},$$

which is decreasing in q. For $f(1;,q)$ and $1 - F(1;q)$,

$$\frac{f_q(1;q)}{f_q(2;q) + f_q(3;q)} = \frac{1}{2 - 2q + 4q} = \frac{1}{2 + 2q},$$

which is decreasing in $q \in [0,1]$. Hence TSCP holds. □

1.7 Conclusion

I show with a stylized static model that when the number of referrals follows a binomial distribution and the effort space is single-dimensional and compact, the highest-threshold contracts would be the most cost-effective way of inducing customers to exert referral efforts. In a number of important extensions, the optimal contract can deviate from the highest-threshold contract, but the paper shows that the firm still has an incentive to include a threshold contract in its referral programme. For managerial implications, this paper provides new and interpretable insights about how the use or inclusion of threshold contracts can help a firm make its referral programme more cost-effective.

One limitation of this paper is that efforts have perfect spillovers on the conversion probability of each friend. This assumption makes it easier to have the customer's action space be one-dimensional, which helps to implement the first-order approach, but the assumption does eliminate the likely factor that customer referral efforts could have different effects on each referrable friend. For the sake of giving a more comprehensive understanding of the effort-incentivizing part of referral programmes, it would be good for future works to have a model where the referral efforts have only partial spillovers on each referrable friend.

1.A Auxiliary Results

1.A.1 Results about CDFC

In Grossman and Hart (1983) and Rogerson (1985), one condition used to make the first-order approach valid is the convex distributional function condition, which is defined here.

Definition 3 (CDFC)**.** The effort induced distribution $F(\cdot;q)$ satisfies the convexity of distributional function condition if $F(n;h(x))$ is convex in x, where $h(\cdot)$ is the inverse function of c with domain $[0, c(1)]$.

The following result shows that the effort-induced binomial distribution does not satisfy CDFC for every convex $c(\cdot)$.

Proposition 16. *If $N \geq 4$, the effort induced distribution is binomial, and the cost function $c(\cdot)$ satisfies A2, then CDFC does not hold.*

Proof. Pick $0 < n < N$ such that $2n < N - 1$. By the property of the binomial distribution,

$$\frac{\partial F(n; h(x))}{\partial x} = F_q(n; h(x))h'(x) \propto (h(x))^n(1 - h(x))^{N-n-1}h'(x),$$

and hence

$$\frac{\partial^2 F(n; h(x))}{\partial x^2} \propto \left[n(h)^{n-1}(1 - h)^{N-n-1} - (N - n - 1)(h)^n(1 - h)^{N-n-2}\right](h')^2 + (h)^n(1 - h)^{N-n-1}h''.$$

Since $c''(q) > 0$ for $q \in (0, 1)$, $h''(x) < 0$ for $x \in (0, c(1))$. In addition, the first term in the right-hand side is negative for $\frac{h(x)}{1-h(x)} \geq \frac{n}{N-n-1}$, where $\frac{n}{N-n-1} < 1$. Therefore, $F(n; h(x))$ is not convex in x and hence CDFC does not hold. □

1.A.2 Result about IMCP

The following result shows that if IMCP holds, then for every target interior effort level, there exists an N-threshold contract that induces the effort level.

Proposition 17 (N-threshold effort-inducing). *Assume IMCP and A1 hold. For every fixed $\theta \in (0, 1)$ and $1 \leq n \leq N$, there exists a unique N-threshold contract that induces θ.*

Proof. Let t be the N-threshold contract such that $v_q(t; \theta) = c'(\theta)$. The claim is that t induces θ. To show this, note that at θ,

$$v_q(t; \theta) = t_N f_q(N; \theta) = c'(\theta).$$

Since $f_q(N; q) > 0$ for every $q \in (0, 1)$, by IMCP, $v_q(t; q) > c'(q)$ for $q \in (0, \theta)$ and $v_q(t; q) < c'(q)$ for $q \in (\theta, 1)$. Thus in t, to the customer, θ is the unique optimal effort level. Therefore, t induces θ. □

1.B Proofs

1.B.1 Proof of Proposition 2

Proof. The proof of the existence of a solution to (1.2) and the strict positiveness of the customer' effort level is referred to Theorem 2. Assume the customer chooses $\theta \in (0, 1]$ in a contract. If $\theta = 1$, then $\lim_{q\to 1} c(q) < \infty$ for such a contract to exist. Since $f_q(N; 1) > 0$, there exists an N-threshold contract that induces 1 if $\lim_{q\to 1} c'(q) < \infty$. Let t be an N-threshold contract inducing θ, and b be another contract in which the customer chooses θ. If b is an N-threshold contract, then since $v_q(t; \theta) = c'(\theta)$, $b_N \geq t_N$ and thus $v(b; 1) \geq v(t; 1)$. Assume instead that b is not an N-threshold contract. At θ, $v_q(b; \theta) \geq c'(\theta) = v_q(t; \theta)$, which implies

$$\sum_{n=1}^{N} b_n \binom{N}{n} \left[n\theta^{n-1}(1 - \theta)^{N-n} - (N - n)\theta^n(1 - \theta)^{N-n-1}\right] \geq t_N N\theta^{N-1}.$$

Multiply both sides by $\frac{\theta}{N}$ and rearrange the terms to get

$$\sum_{n=1}^{N} b_n \binom{N}{n} \frac{n}{N} \theta^n (1-\theta)^{N-1} - t_N \theta^N \geq \sum_{n=1}^{N} b_n \binom{N}{n} \frac{N-n}{N} \theta^{n+1} (1-\theta)^{N-n-1} > 0,$$

where the last inequality is strict because $b_n > 0$ for some $n < N$ for b is not an N-threshold contract. Since

$$\begin{aligned} v(b;\theta,N) - v(t;\theta,N) &= \sum_{n=1}^{N} b_n \binom{N}{n} \theta^n (1-\theta)^{N-1} - t_N \theta^N \\ &> \sum_{n=1}^{N} b_n \binom{N}{n} \frac{n}{N} \theta^n (1-\theta)^{N-1} - t_N \theta^N > 0 \end{aligned}$$

the N-threshold contract has a strictly lower expected cost to the firm. Therefore, every optimal contract must be an N-threshold contract. □

1.B.2 Proof of Proposition 3

Proof. For b to incentivize θ,

$$v_q(b;\theta) = -\sum_{n=1}^{N} \Delta b_n F(n-1;\theta) = c'(\theta).$$

Additionally, since $F_q(n-1;q) < 0$ in $q \in (0,1)$ and $v_q(t(n);\theta) = -t_n(n)F(n-1;\theta) = c'(\theta)$ for $1 \leq n \leq N$, $0 \leq \Delta b_n \leq t_n(n)$. For the same n, since $v_q(t(n);\theta) = c'(\theta)$, $t_n(n) > 0$. Let $\lambda_n = \frac{\Delta b_n}{t_n(n)} \in [0,1]$ for each $1 \leq n \leq N$. By construction, $\lambda_n \in [0,1]$ for $1 \leq n \leq N$ and $b = \sum_{n=1}^{n} \lambda_n t(n)$. Additionally, since $v_q(b;\theta) = v_q(b(n);\theta)$ for each $1 \leq n \leq N$, by the linearity of $v_q(\cdot;q)$ for $q \in [0,1]$, $\sum_{n=1}^{N} \lambda_n = 1$ and hence b is in the convex hull of $\{t(n)\}_{n=1}^{N}$.

Lastly, for each $1 \leq n \leq N$, θ is the unique optimal effort level to the customer in $t(n)$. By the linearity of $v(\cdot;q)$ for $q \in [0,1]$, θ must be the unique optimal effort level to the customer in b, i.e., b induces θ. □

1.B.3 Proof of Proposition 4

Proof. The following lemma is very useful in proving the proposition.

Lemma 2 (Partial order on contracts)**.** *Fix $\theta \in (0,1)$. Let b and b' be two different contracts such that both induce θ and $b_0 = b'_0$. If there exists some n^* such that $\Delta b_n - \Delta b'_n \geq 0$ for $n \geq n^*$ and $\Delta b_n - \Delta b'_n \leq 0$ for $n \leq n^*$, then b has a strictly lower expected cost to the firm.*

Proof of lemma. Since both contracts induce θ and $c' > 0$ on $(0,1)$, we get $v_q(b;q) > 0$ and $v_q(b';q) > 0$ for $q \in (0,\theta)$. Note that $v_q(b;q,N) \leq v_q(b';q,N)$ for some $q \in (0,1)$ is equivalent to

$$-\sum_{n=1}^{N} \Delta b_n F_q(n-1;q) \leq -\sum_{n=1}^{N} \Delta b'_n F_q(n-1;q),$$

which is equivalent to

$$-\sum_{n=n^*}^{N} (\Delta b_n - \Delta b'_n) F_q(n-1;q) \leq -\sum_{1}^{n^*-1} (\Delta b'_n - \Delta b_n) F_q(n-1;q). \tag{1.11}$$

The inequality binds at $q = \theta$. Since strict MLRP holds, $-F_q(n-1;q) > 0$. In addition, as $\Delta b_n - \Delta b'_n \geq 0$ for $n \geq n^*$ and $\Delta b_n - \Delta b'_n \leq 0$ for $n \leq n^*$, every term (including the preceding minus sign) in the summations on both sides are non-negative. If either side sums up to 0, then either $b = b'$ or the inequality does not bind at $q = \theta$, a contradiction. Assume instead that both sides are strictly positive.

Note that $F_q(n;q) < 0$ for $0 \leq n < N$. If for every $0 \leq n < N$, $\frac{F_q(n-1;q)}{F_q(n;q)}$ is strictly decreasing in $q \in (0,1)$, which is the strict TSCP introduced in the main text, then (1.11) would hold strictly for every $q \in (0,\theta)$. Since $b_0 = b'_0 = 0$, $v(b;\theta) = \int_0^\theta v_q(b;\theta)$ and $v(b';\theta) = \int_0^\theta v_q(b';\theta)$. Thus $v(b;\theta) < v(b';\theta)$ and the result is proved. Therefore, to complete the proof, it suffices to show that the binomial distribution satisfy TSCP. For each $0 \leq n \leq N$,

$$\begin{aligned}
F_q(n;q) = \sum_{i=0}^{n} f_q(i;q) &= \sum_{i=0}^{n} \binom{N}{i} \left[i q^{i-1}(1-q)^{N-i} - (N-i) q^i (1-q)^{N-i-1} \right] \\
&= \sum_{i=0}^{n-1} \left[\binom{N}{i+1}(i+1) - \binom{N}{i}(N-i) \right] q^i (1-q)^{N-i-1} \\
&= -\binom{N}{n}(N-n) q^n (1-q)^{N-n-1}.
\end{aligned} \tag{1.12}$$

Thus for every $n_1 < n_2$,

$$\frac{F_q(n_1;q)}{F_q(n_2;q)} \propto \left(\frac{1-q}{q} \right)^{n_2 - n_1},$$

which is strictly decreasing in q on $(0,1)$. Therefore, the effort-induced binomial distribution satisfies TSCP and hence $v(b;\theta) < v(b';\theta)$. The proof of the lemma is complete. □

Assume $n > n'$ and pick any integer n^* from n to n'. For $n \leq n^*$, $\Delta t_n - \Delta t'_n \leq 0$; for $n \geq n^*$, $\Delta t_n - \Delta t'_n \geq 0$. Thus t has a strictly lower expected cost by Lemma 2. □

1.B.4 Proof of Proposition 5

Proof. If $n = N$, by the optimality of N-threshold contracts established in Proposition 2, $q_N^* = 1$. Fix $n < N$. Let $b(\theta)$ be the linear contract with incremental payment $\ell_\theta > 0$ that induces θ and let $t(n;\theta)$ be the n-threshold contract that induces the same effort level. Since $v_q(t(1);q) < 0$ for $q \in (0,1)$, $q_1^* = 0$. Assume instead $1 \le n < N$. At θ, the first-order conditions of the two contracts imply $v_q(b(\theta;\theta)) = v_q(t(n;\theta);\theta)$, which is

$$\ell_\theta N = -t_n(n;\theta)F_q(n-1;\theta).$$

Since $v_q(t(n;\theta);q)$ is strictly increasing for $\frac{q}{1-q} < \frac{n-1}{N-n}$ and strictly decreasing beyond the threshold. Therefore, $t_n(n;\theta)/\ell_\theta$ is decreasing in θ for $\frac{\theta}{1-\theta} < \frac{n-1}{N-n}$ and increasing after the threshold. If $\frac{\theta}{1-\theta} < \frac{n-1}{N-n}$, then since $v_q(t(n;\theta);q)$ is increasing and $v_q(b(\theta);q)$ is constant for $q < \theta$, $v(b(\theta);\theta) > v(t(n;\theta);\theta)$. Now define

$$D(\theta) = \frac{v(t(n;\theta);\theta) - v(b(\theta);\theta)}{\ell_\theta} = \frac{t_n(n;\theta)}{\ell_\theta}\int_0^\theta n\binom{N}{n}q^{n-1}(1-q)^{N-n}dq - N\theta,$$

where the first part of the right-hand side comes from (1.12) and the fact that $f(n;0) = 0$ for all $n > 0$. Hence

$$\begin{aligned} D'(\theta) &= \frac{\partial}{\partial\theta}\left(\frac{t_n(n;\theta)}{\ell_\theta}\right)\int_0^\theta v_q(t(n;\theta);q)dq + \frac{v_q(t(n;\theta);\theta)}{\ell_\theta} - N. \\ &= \frac{\partial}{\partial\theta}\left(\frac{t_n(n;\theta)}{\ell_\theta}\right)\int_0^\theta v_q(t(n;\theta);q)dq, \end{aligned}$$

where the last equality comes from the first-order condition of the two contracts at θ. Hence $D'(\theta)$ has the same sign with $\frac{\partial}{\partial\theta}(t_n(n;\theta)/\ell_\theta)$. Therefore, $D(\theta) < 0$ and is decreasing for $\frac{\theta}{1-\theta} < \frac{n-1}{N-n}$ and increasing otherwise. Since $\ell_\theta > 0$ and is increasing in θ, if $D(\theta) > 0$, then $D'(\theta) > 0$ and $D(q) > 0$ for all $q > \theta$. Thus, there exists some $q_n^* \in [0,1]$ such that $v(t(n;\theta);\theta) \le v(b(\theta);\theta)$ if and only if $\theta \le q_n^*$.

Fix $1 \le n < N$. To show that $q_{n+1}^* > q_n^*$ if $q_n^* < 1$, note that by Proposition 4, the $(n+1)$-threshold contract that induces θ has a strictly lower expected cost than does the n-threshold contract. Therefore, $v(t(n+1;q_n^*);q_n^*) < v(t(n;q_n^*);q_n^*) \le v(b(\theta);q_n^*)$ and hence $q_{n+1}^* > q_n^*$ if $q_n^* < 1$. □

1.B.5 Proof of Proposition 6

Proof. Since a strictly profitable contract with respect to (1.3) exists, in every solution to (1.3), the firm gets a strictly positive profit, and hence in every solution to (1.3), the effort level must be strictly positive.

Let t^* be the N-threshold contract that incentivizes q_{FB}. Define $\underline{r} = v(t^*;q_{FB}) - c(q_{FB})$ and $\bar{r} = \pi N q_{FB} - c(q_{FB})$.

$r \leq \underline{r}$ Assume $r \leq \underline{r}$. Let t be the N-threshold contract ignoring the participation constraint. Let $q(\cdot)$ map a threshold contract to the incentivized effort level. By Remark 1, $q(t) < q_{FB}$. If the participation constraint does not hold, then raising t_N will incentivizing the customer to work more as well as improving the customer's payoff. Raise t_N until the customer's participation constraint binds and denote the new contract by t'. Now let $b \neq t'$ be a contract that is feasible in (1.3) and incentivizes an effort level less than $q(t')$. From b to t', the rise of expected reward is capped above by $c(q(t')) - c(q(t))$ and the firm's referral revenue increases by $\pi N(q(t') - q(t))$. Since $q(t') \leq q_{FB}$ by the assumption of r, the increase in revenue is more than the increase in expected reward. Therefore, t' is better than any contract that is feasible in (1.3) and incentivizes an effort level less than $q(t')$.

Now suppose b is feasible in (1.3) and incentivizes an effort level more than $q(t')$, there exists an N-threshold contract that incentivizes the effort level, i.e., the participation constraint holds. Since the N-threshold contract is uniquely cost-efficiently, b is not optimal. Therefore, t' is optimal when $r < \underline{r}$.

$r \geq \bar{r}$ If $r \geq \bar{r}$, then the firm's first-best profit is zero, and hence it is optimal not to have a referral programme.

$r \in (\underline{r}, \bar{r})$ Assume $r \in (\underline{r}, \bar{r})$. Let t be the N-threshold contract incentivizing q_{FB} and ℓ be the linear contract incentivizing q_{FB}. To be precise, $\Delta \ell = \pi$, i.e., the customer gets all the referral revenue as referral reward. By construction, the participation constraint does not hold in t and the constraint holds strictly in ℓ. Therefore, there exists some $\lambda \in (0, 1)$ such that $v(\lambda t + (1-\lambda)\ell; q_{FB}) - c(q_{FB}) = r$. It suffices to show that the mixture incentivizes q_{FB}. To see this, note that

$$v_q(t; q) - c'(q) = \begin{cases} \leq 0 & q \leq q_{FB} \\ > 0 & q > q_{FB} \end{cases}, \quad v_q(\ell; q) - c'(q) = \begin{cases} \leq 0 & q \leq q_{FB} \\ > 0 & q > q_{FB} \end{cases}.$$

Since $v_q(\cdot; q)$ is linear in the contract, the single-crossing property continues to hold for $\lambda t + (1-\lambda)\ell$. Therefore, the combined contract cost-efficiently incentivizes q_{FB} and the participation constraint binds. □

1.B.6 Proof of Theorem 1

Proof. The proof uses a similar Lagrangian approach to the one used in Innes (1990). Consider the firm's relaxed profit-maximization problem

$$\max_{\substack{b \in \mathbb{R}^{N+1} \\ q \in [0,1]}} \quad \pi N q - v(b; q) \tag{1.13}$$

$$\text{s.t.} \quad v(b; q) \geq c(q) \tag{1.13a}$$

$$v_q(b; q) \geq c'(q) \quad \forall q \in [0, 1] \tag{1.13b}$$

$$0 \leq b_n \leq \pi n \qquad \forall 0 \leq n \leq N. \tag{1.13c}$$

This problem is a relaxed problem of (1.4) because (1.4b) implies (1.13b). The relaxed problem has a solution because the objective function is continuous and the feasible set is non-empty and compact. Assume a contract that is strictly profitable with respect to (1.4) exists. Since the firm has zero revenue (before referral rewards) if $q = 0$, in every solution to (1.13), $q > 0$.

Let (b, q) be a solution to (1.13). The Lagrangian function of (1.13) associated with (b, q) is

$$L = \pi N q - v(b; q) - \lambda[v(b; q) - c(q)] - \mu[v_q(b; q) - c'(q)] + \sum_{n=0}^{N} \eta_n b_n - \psi[b_n - \pi n].$$

The Lagrangian function's first-order condition with respect to b_n is

$$L_{b_n} = -f(n; q) + \lambda f(n; q) + \mu f_q(n; q) + \eta_n - \psi_n.$$

Assume $\mu > 0$ and $q \in (0, 1)$. Since $\frac{f_q(n;q)}{f(n;q)}$ is strictly increasing in n, if $b_n > 0$ and $b_{n+1} < \pi(n = 1)$, $L_{b_{n+1}} > 0$, a contradiction. Hence if $b_n > 0$, then $b_{n+1} = \pi(n + 1)$.

Assume $\mu > 0$ and $q = 1$. Since $\mu > 0$, the customer's first-order condition binds. Since $f(N; 1) = 1$, $f_q(n; q) \leq 0$ for each $n < N$. Therefore, if the customer chooses 1 in b, then there exists an N-threshold contract t inducing 1. By the proof of Proposition 2, $v(b; 1) > v(t; 1)$, contradicting the optimality of b.

Assume $\mu = 0$. Since $b \neq 0$, it is necessary that $\lambda > 0$ and hence the customer's participation constraint binds. If $b_n > 0$ but $b_{n+1} < \pi(n + 1)$, by lowering b_n and adjusting b_{n+1} simultaneously by an appropriate ratio, the customer's participation constraint binds, and the relaxed first-order condition holds. However, this implies the existence of a non-decreasing contract inducing a strictly positive effort level yet gets zero payoffs, which is a contradiction because by Proposition 3, the first-order condition uniquely determines the customer's optimal effort level. Therefore, $\mu > 0$.

It has been shown that (b, q) to (1.13) must have the property that $b_n > 0$ implies $b_{n+1} = \pi(n+1)$. Since $\mu > 0$ and the contract is non-decreasing, the customer optimally choose q in such a contract, and hence (b, q) is a solution to the actual optimization problem (1.4). □

1.B.7 Proof of Proposition 8

Proof. The relaxed optimization problem of (1.5) is

$$\max_{\substack{b \in \mathbb{R}^{N+1} \\ q \in [0,1]}} \quad \pi N q - \sum_{n=0}^{N} b_n f(n; q) \tag{1.14}$$

$$\text{s.t.} \quad \sum_{n=0}^{N} V(b_n) f(n; q) - c(q) \geq 0 \tag{1.14a}$$

$$\sum_{n=0}^{n} V(b_n) f_q(n; q) \geq c'(q) \tag{1.14b}$$

$$b_n \geq 0, \quad 0 \leq n \leq N \tag{1.14c}$$

Let (b, q) be a solution to (1.14). The Lagrangian associated with (b, q) is

$$L = \pi N q - \sum_{n=0}^{N} b_n f(n; q) + \lambda \left[\sum_{n=0}^{N} V(b_n) f(n; q) - c(q) \right] + \mu \left[\sum_{n=0}^{n} V(b_n) f_q(n; q) - c'(q) \right] + \sum_{n=0}^{N} \eta_n b_n, \quad (1.15)$$

where $\lambda \geq 0$, $\mu \geq 0$, and $\eta_n \geq 0$ for every n are the Lagrangian multipliers associated with b. For each n,

$$L_{b_n} = -f(n; q) + \lambda V'(b_n) f(n; q) + \mu V'(b_n) f_q(n; q) + \eta_n. \quad (1.16)$$

If $\mu = 0$, then $\lambda > 0$ and the customer's participation constraint binds. MLRP implies that $\frac{f_q(n;q)}{f(n;q)}$ is non-decreasing in q. Therefore, whenever $b_n > 0$ and $n < N$, by lowering b_n and raising b_N by an appropriate ratio, the customer's participation constraint stays bound and the customer's relaxed first-order condition still holds. However, this implies the existence of an N-threshold contract that induces at least q and the customer's participation constraint binds, which is a contradiction since the customer gets strictly payoff from a non-decreasing contract inducing the customer to choose a strictly positive effort level. Therefore, $\mu > 0$, i.e., at (b, q), the customer's first-order condition binds.

If $b_n > 0$ but $f_q(n; q) < 0$, then $L_{b_n} < 0$, a contradiction. Note that $f_q(n; q) < 0$ if and only if $q > n/N$. Therefore, $b_n = 0$ if $q > n/N$. To complete the proof, it is left to show that (b, q) as a solution to (1.14) also solves (1.5). By MLRP, for $b_n = b_{n+1}$, $L_{b_{n+1}} \geq L_{b_n}$, implying that the solution must be non-decreasing. With a similar reasoning to Proposition 3, the first-order approach is valid for (b, q) and hence (b, q) solves (1.5). □

1.B.8 Proof of Proposition 9

Proof. The proof uses the conditions in Grossman and Hart (1983) to prove the convexity of optimal contracts. The firm solves the same optimization problem (1.5) with the same relaxed problem (1.14). With a similar reasoning to that in the proof of Proposition 8, $\mu > 0$ and $\lambda = 0$ in (1.15).

Since strict MLRP holds for $q \in (0, 1)$, in (1.16), for $b_n = b_{n'}$ with $n' > n$, if $L_{b_n} \geq 0$, then $L_{b_{n'}} > 0$. Thus in every solution, if $b_n > 0$, then $b_{n+1} > b_n$. In addition, for $b_n > 0$ in a solution, the first-order condition is

$$\frac{1}{V'(b_n)} = \mu \frac{f_q(n; q)}{f(n; q)},$$

and thus

$$\frac{1}{V'(b_{n+1})} - \frac{1}{V'(b_n)} = \mu \left[\frac{f_q(n+1; q)}{f(n+1; q)} - \frac{f_q(n; q)}{f(n; q)} \right].$$

With $\mu > 0$ and $1/V'(\cdot)$ concave, if it can be shown that $\frac{f_q(n+1;q)}{f(n+1;q)} - \frac{f_q(n;q)}{f(n;q)}$ is non-decreasing in n, which is a condition introduced in Grossman and Hart (1983), then $b_{n+2} - b_n \geq b_{n+1} - b_n \geq 0$.

Given the binomial distribution,

$$\frac{f_q(n;q)}{f(n;q)} = \frac{n}{q} - \frac{N-n}{1-q},$$

and hence $\frac{f_q(n+1;q)}{f(n+1;q)} - \frac{f_q(n;q)}{f(n;q)} = \frac{1}{q} + \frac{1}{1-q}$, which is constant and hence non-decreasing in n. □

1.B.9 Proof of Proposition 10

Proof. Let $\ell \in \mathbb{R}^{N+1}$ be an optimal linear contract with respect to (1.7), with $\beta = \Delta\ell_1$ being the constant incremental reward in the linear contract. If (1.7) is zero in ℓ, then the proposition is trivially true. Now assume instead that (1.7) is strictly positive in ℓ. Since the objective function is zero with $q = 0$ or the number of messages being zero, in ℓ, the customer chooses an effort level $q > 0$ and sends a positive number of messages at optimum.

For the customer to send a positive number of messages, it is necessary that $\beta q \geq s$. Since (1.7) is strictly positive at optimum, $\beta < \pi$ and hence the customer must send N referral messages at the firm's optimum. It thus remains to prove the existence of a contract that is a combination of a linear contract and an N-threshold contract, feasible in (1.7), and improves (1.7).

By treating the communication cost of sending N messages in the customer's reserve utility, Proposition 6 implies that there exists a contract $b \in \mathbb{R}^{N+1}$ that is a combination of a linear contract and an N-threshold contract, such that conditional on that the customer has an incentive to send N messages, b improves (1.7) over ℓ. Therefore, it suffices to show that in b, the customer has an incentive to send N messages.

Let $b = \ell^* + t$, where $\ell^* \in \mathbb{R}^{N+1}$ is the linear part of b and $t \in \mathbb{R}^{N+1}$ the N-threshold part. Since the customer's participation constraint holds in b, the customer has no incentive to send no messages. Now assume that it is optimal for the customer to send $\tilde{N} < N$ messages and exert effort level $q > 0$. By the construction of b, for $\tilde{N} < N$, the customer is in a linear contract, and hence $q\Delta\ell_1^* \geq s$, and thus $\tilde{N} = N - 1$ is optimal to the customer. However, if the customer sends N messages instead of $N-1$ with the same effort level q, then the extra expected reward is $q\Delta\ell_1^* + f(N;q,N) > s$, and hence the customer is strictly better off by sending N messages instead, a contradiction. Hence in b, sending N messages is uniquely optimal and hence b is feasible in (1.7). Therefore, b improves (1.7) over ℓ. □

1.B.10 Proof of Proposition 11

Proof. If there exists no menu that is strictly profitable with respect to (1.9), then the zero menu is optimal. If an optimal menu exists and is strictly profitable, then at least one type must choose a strictly positive effort level.

It can be shown both types choose strictly positive effort levels. Let (b^1, b^2) be an optimal menu that is strictly profitable. To arrive at a contradiction, first assume that only the low type chooses a strictly positive effort level. The firm's profit from the high type is non-positive by the limited liability of customer assumption. If $V^*(b^1; N_2) \leq 0$, then the firm can offer the high type the optimal contract $t \in \mathbb{R}^{N_2+1}$ when there is only the high type. By Proposition 2, t is an N_2-threshold contract. Therefore, (b^1, t) is incentive-compatible because $V^*(t; N_1) = 0$ and $V^*(t; N_2) > 0$. Therefore, $\Pi(b^1, t) > \Pi(b^1, b^2)$, a contradiction. If $V^*(b^1; N_2) > 0$, then $b_0^2 > 0$. It is without loss of generality to assume that $b^2 = V^*(b^1; N_1)$ and $b_n^2 = 0$ for $n > 0$. Assume that the firm uses an N_2-threshold contract t to induce the high type to choose some small $\theta_2 > 0$. To make the new menu incentive compatible, the firm also needs to give the high type an unconditional reward of r to make the high type's IC constraint bind. Since the high type's utility is increasing in t_{N_2}, $r < b_0^2$, and hence $V^*(t + r; N_1) < V^*(b^1; N_1)$. The high type's expected rewards increase by $c(\theta_2)$, whereas the firm's revenue increase by $\pi N_2 \theta_2$. Since $c'(q) < \pi N_2$ for small q, $\pi N_2 \theta_2 > c(\theta_2)$, and thus $\Pi(b^1; t+r) > \Pi(b^1, b^2)$, contradicting the optimality of (b^1, b^2). Therefore, if the low type chooses a strictly positive effort level in an optimal contract, then the high type also chooses a strictly positive effort level.

If only the high type exerts efforts in an optimal menu, then by Proposition 2, it is optimal to have b^2 be an N_2-threshold contract. Assume the high type chooses $\theta > 0$ in b^2. From the proof of Proposition 2, $v(b^2; \theta, N_2) > c(\theta)$. However, this implies that if b^1 is set to an N_1-threshold contract with $b_{N_1}^1 > 0$ being small, (b^1, b^2) is incentive compatible and the low type chooses a strictly positive effort level. In addition, for small $b_{N_1}^1 > 0$, the firm gets a strictly positive profit from the low type. Therefore, in every optimal contract, both types exerts strictly positive efforts.

It remains to show that (1.9) has a solution. To see this, note that the limited liability of customer assumption bounds the objective function above. Additionally, since the objective function is continuous and the feasible set of (1.9) is closed, a maximizer exists. □

1.B.11 Proof of Proposition 12

Proof. Let (b^1, b^2) be an optimal menu and assume b^2 is not an N_2-threshold contract. Let $q_2' = q^*(b^2; N_2)$ and t^2 be the N_2-threshold contract inducing the high type to choose q_2'. By the proof of Proposition 2, $v(t^2; q_2', N_2) < v(b^2; q_2', N_2)$.

If $V^*(b^1; N_2) \geq V^*(t^l; N_2)$, let $t^1 = t^l$; otherwise, let $t^1 = b^1$. If $V^*(t^2; N_2) \geq V^*(t^1; N_2)$, then (t^1, t^2) is incentive-compatible. When $t^1 = t^l$, the firm gets a weakly higher profit from the low type since t^l is optimal when there is only the low type. Additionally, since the high type exerts the same effort level yet at a strictly lower expected rewards in t^2 than in b^2, the firm gets a strictly higher profit from the high type in (t^1, t^2). Therefore, $\Pi(b^1, b^2) < \Pi(t^1, t^2)$, a contradiction.

Assume instead that $V^*(t^2; N_2) < V^*(t^1; N_2)$, i.e., (t^1, t^2) is not incentive-compatible. In this case, redefine t^2 by raising $t_{N_2}^2$ so that $V^*(t^2; N_2) = V^*(t^1; N_2)$. Let $q_2 = q^*(t^2; N_2)$. By the choice of t^1, since $V^*(t^1; N_2) \leq V^*(b^1; N_2)$ and $V^*(t^2; N_2)$ is strictly increasing in $t_{N_2}^2$, $q_2 \leq q_{FB}(N_2)$. By the choice of (t^1, t^2),

$$V^*(t^2; N_2) = V^*(t^1; N_2) \leq V^*(b^1; N_2) \leq V^*(b^2; N_2),$$

which means

$$c(q_2) - c(q_2') \geq v(t^2; q_2, N_2) - v(b^2; q_2', N_2).$$

Since $q_2 \leq q_{FB}(N_2)$ and $q_2 > q_2'$,

$$\pi N_2 (q_2 - q_2') > c(q_2) - c(q_2') \geq v(t^2; q_2, N_2) - v(b^2; q_2', N_2).$$

Thus the firm gets a strictly higher profit from the high type and $\Pi(t^1, t^2) > \Pi(b^1, b^2)$, contradicting the optimality of (b^1, b^2). Therefore, if (b^1, b^2) is optimal, then b^2 is an N_2-threshold contract. □

1.B.12 Proof of Proposition 13

Proof. If there exists no contract that is strictly profitable with respect to (1.10) exists, then the zero contract is optimal. Assume instead that such a strictly profitable contract exists. In that case, if an optimal contract exists, then the firm's profits would be strictly positive, and at least one type chooses a strictly positive effort level. Since referral contracts are now assumed to be non-decreasing, if the low type chooses a strictly positive effort level, then by exerting the same effort level, the high type would get at least the payoff of the low type. Therefore, the high type always chooses a strictly positive effort level.

Because of the limited liability of customer, the firm's profits cannot exceed $\pi(N_1+N_2)$. Therefore, let $\overline{\Pi} = \sup\{\Pi(b, b)\} \in (0, \infty)$ and find a sequence of contracts $(b^i)_{i=1}^{\infty}$, where each $b^i \in \mathbb{R}^{N_2+1}$ such that $\Pi(b^i, b^i) \leq \Pi(b^{i+1}, b^{i+1})$ for every i and $\lim_{i\to\infty} \Pi(b^i, b^i) = \overline{\Pi}$. Since $[0, 1]^2$ is compact, it can be assumed that $\theta_1^i = q^*(b^i; N_1)$ and $\theta_2^i = q^*(b^i; N_2)$ converge to some θ_1 and θ_2.

It remains to show that (1.10) admits a solution. The objective function is bounded above thanks to the limited liability of customer assumption. Additionally, since the feasible set is closed, (1.10) has a solution. □

1.B.13 Proof of Lemma 1

Proof. If there exists $\tilde{b} \in \mathbb{R}^{N_2+1}$ such that $\tilde{b}_0 > 0$ and the two types choose (θ_1, θ_2), define $b = \tilde{b} - \tilde{b}_0$, which is non-decreasing and non-negative contract since $\tilde{b}$ is non-decreasing and non-negative. By Proposition 3, the customer would still choose θ_1 if given contract b. At every $q \in (0, 1)$,

$$v(\tilde{b}; q, N_1) = \tilde{b}_0 + \sum_{n=1}^{N_2} \Delta b_n \left[1 - F(n-1; q, N_1)\right] = v(b; q; N_1) - \tilde{b}_0,$$

and thus b pays the low type lower expected rewards. Similarly, it can be shown that the high type would choose θ_2 under b and b has a lower expected cost to the firm. Therefore, given (θ_1, θ_2), every cost-efficient contract must have $b_0 = 0$. For the rest of the proof, set $b_0 = 0$.

If $\theta_1 = 0$, then by Proposition 2, the N_2-threshold inducing θ_2 by the high type is cost-efficient. Assume $(\theta_1, \theta_2) \in [0, 1]^2$. Consider the following relaxed cost-minimization problem to the firm:

$$\min_{b\in\mathbb{R}^{N_2}} \quad \alpha v(b; \theta_1, N_1) + (1-\alpha) v(b; \theta_2, N_2) \tag{1.17}$$

$$\text{s.t.} \quad v_q(b; \theta_i, N_i) = c'(\theta_i), \quad 1 \leq i \leq 2 \tag{1.17a}$$

$$0 \leq b_1 \leq b_2 \leq \cdots \leq b_{N_2} \tag{1.17b}$$

Problem (1.17) has the same solution to the actual cost-minimization part of (1.10), because by Proposition 3, the first-order condition is sufficient for the effort-choice constraints with non-decreasing contracts. Since (1.17) is a linear programming problem with linear constraints, there exists a minimizer that is an extreme point of the feasible set of (1.17).

In (1.17), each extreme point is found by binding N_2 linearly independent constraints. Since the zero contract is not a solution, for each extreme point, either exactly $N-1$ or $N-2$ of the inequality constraints bind, which is a contract with two thresholds. Since the solution to (1.17) also solves the cost-minimization part of (1.10), (1.10) admits a solution that has a contract with two thresholds. □

1.B.14 Proof of Proposition 14

Proof. The following lemma is useful in proving the proposition.

Lemma 3 (Highest-threshold for high type)**.** *Fix $(\theta_1, \theta_2) \in (0,1]^2$. Assume there exists an (n_1, n_2)-threshold contract such that the two types optimally choose (θ_1, θ_2). If $N_1 < n_2 < N_2$ and $b_{n_2} > b_{n_1}$, then there exists an (n_1, N_2)-threshold contract that induces (θ_1, θ_2) and has strictly lower expected rewards.*

Proof of lemma. Let $\tilde{b} \in \mathbb{R}^{N_2}$ be an (n_1, n_2)-threshold contract that induces (θ_1, θ_2). Let $\hat{b} \in \mathbb{R}^{N_2}$ be an n_1-threshold contract such that $\hat{b}_n = \tilde{b}$ for $n \leq n_1$ and $\hat{b}_n = \hat{b}_{n_1}$ for $n \geq n_1$. As $\tilde{b}_{n_2} > \tilde{b}_{n_1}$, $\tilde{b}$ is non-decreasing, and $F(n; q, N_2)$ is strictly decreasing in q for $q \in (0,1)$, $v_q(\hat{b}; \theta_2, N_2) < c'(\theta_2)$. Since $f_q(N; q, N_2) > 0$ for $q \in (0,1)$, there exists an (n_1, N_2)-threshold contract b that induces (θ_1, θ_2) and $b_n = \tilde{b}$ for $1 \leq n \leq n_1$. It can be shown that b strictly improves the firm's cost. At θ_2, $v_q(\tilde{b}; \theta_2, N_2) = v_q(b; \theta_2, N_2)$, which implies

$$\binom{N_2}{n_1}\tilde{b}_{n_1} n_1 \theta_2^{n_1-1}(1-\theta_2)^{N_2-n_1} + \binom{N_2}{n_2}\left(\tilde{b}_{n_2} - \tilde{b}_{n_1}\right) n_2 \theta_2^{n_2-1}(1-\theta_2)^{N_2-n_2}$$
$$= \binom{N_2}{n_1}\tilde{b}_{n_1} n_1 \theta_2^{n_1-1}(1-\theta_2)^{N_2-n_1} + \left(b_N - \tilde{b}_{n_1}\right) N_2 \theta_2^{N_2-1}.$$

The equation can be simplified to

$$\binom{N_2}{n_2}\left(\tilde{b}_{n_2} - \tilde{b}_{n_1}\right) n_2 = \left(b_N - \tilde{b}_{n_1}\right) N_2 \left(\frac{\theta_2}{1-\theta_2}\right)^{N_2-n_2}.$$

Since the right-hand side is strictly increasing in θ_2,

$$\left(\tilde{b}_{n_2} - \tilde{b}_{n_1}\right) n_2 > \left(b_N - \tilde{b}_{n_1}\right) N_2 \left(\frac{q}{1-q}\right)^{N_2-n_2}$$

for every $q \in (0, \theta)$. Hence $v(\tilde{b}; \theta_2, N_2) > v(b; \theta_2, N_2)$, i.e., b has a strictly lower expected cost to the firm. The proof of the lemma is complete. □

By Proposition 13, the two types exert strictly positive efforts in every optimal contract. Assume the two types choose $(\theta_1, \theta_2) \in (0,1]^2$ in an optimal contract. By Lemma 1, there exists a contract with two thresholds that is cost-efficient to induce (θ_1, θ_2). Let (n_1, n_2) be the two thresholds of this contract. If $n_2 > N_2$, then by Lemma 3, $n_2 = N_2$. □

1.B.15 Proof of Theorem 2

Proof. Since the customer has a limited liability, the objective function in (1.2) is bounded above. By the continuity as assumed in A1 and A2, the constraint set of (1.2) is closed, and hence (1.2) has a solution. Let (b, q) be a solution to (1.2). By A3, $U(b, q) > 0$. Lastly, by A1(c), $q > 0$. □

1.B.16 Proof of Theorem 3

Proof. Note that strict DHRP implies that higher interior efforts strictly first-order stochastically dominate lower efforts, i.e., $F_q(n; q) < 0$ for all $0 \le n < N$ and $q \in (0, 1)$.[17] The following lemma, which establishes the relatively cost-effectiveness of different threshold contracts without consideration of actual effort-choice constraints, is useful.

Lemma 4. *Assume A1 holds. Fix $\theta \in (0, 1)$, n and n' such that $0 \le n' < n < N$. For $\beta > 0$ and $\beta' > 0$ such that $-\beta F_q(n; \theta) = -\beta' F_q(n'; \theta) > 0$, $\beta(1 - F(n; \theta)) < \beta'(1 - F(n'; \theta))$.*

Proof of lemma. Note that

$$\frac{1 - F(n'; q)}{1 - F(n; q)} = \prod_{m=n'}^{n-1} \frac{1 - F(m; q)}{1 - F(m + 1; q)},$$

and each term in the product on the right-hand side is strictly decreasing in $q \in (0, 1)$ by DHRP. Therefore, the left-hand side is strictly decreasing in $q \in (0, 1)$, which implies

$$F_q(n'; q)\,[1 - F(n; q)] < -F_q(n; q)\,[1 - F(n'; q)] \Longrightarrow \frac{-F_q(n'; q)}{-F_q(n; q)} < \frac{1 - F(n'; q)}{1 - F(n; q)},$$

The denominator on the right-hand side of the second inequality is non-zero because $1-F(n; q) > 0$ by A1(b). The second inequality holds because $-F_q(n; q) > 0$. Multiply both sides of the second inequality by β'/β to get

$$1 = \frac{-\beta' F_q(n'; q)}{-\beta F_q(n; q)} < \frac{\beta'\,[1 - F(n'; q)]}{\beta\,[1 - F(n; q)]},$$

and hence $\beta\,[1 - F(n; q)] < \beta'\,[1 - F(n'; q)]$. □

Since A3 holds, by Theorem 2, (1.2) has a solution (b, q) and it is necessary that $q > 0$. Assume

[17] To see this, assume to the contrary that for some $q \in (0, 1)$, $F_q(n; q) \ge 0$ for some $0 \le n < N$. Let n be the smallest n such that $F_q(n; q) \ge 0$. If $n = 0$, then at q, $\frac{f(0;q)}{1-F(0;q)}$ is not strictly decreasing at q, a contradiction. If $n > 0$, then for $\frac{f(n;q)}{1-F(n;q)}$ to be strictly decreasing at q, it is necessary that $f_q(n; q) \le 0$, implying that $F_q(n - 1; q) \ge 0$ at q, a contradiction to the choice of n. Hence $F_q(n; q) < 0$ for all $0 \le n < N$ and $q \in (0, 1)$.

$q < 1$. Given the effort level, consider the following relaxed cost-minimization problem

$$\min_{b \in \mathbb{R}^{N+1}} \quad \pi N q - v(b;q) \tag{1.18}$$

$$\text{s.t.} \quad b_n \sum_{n=0}^{N} f_q(n;q) = c'(q), \tag{1.18a}$$

$$b_n \geq 0, \qquad\qquad 0 \leq n \leq N. \tag{1.18b}$$

The cost-minimization problem is relaxed in the sense that the customer's participation constraint is discarded and the customer's first-order condition is substituted for the actual effort-choice constraint. Given the optimal effort level q, if we can show that (b,q) solves (1.18) and is feasible in (1.2), then (b,q) also solves (1.2).

Let $b \in \mathbb{R}_+^{N+1}$ solve (1.18). Since (1.18) is a linear programming with a linear constraint, there exists a solution that is an extreme point to the constraints. Each extreme point can be found by binding $1+N$ linearly independent constraints. Since $b_n = 0$ for every n cannot solve (1.18), exactly N inequalities are bound in an extreme point that solves (1.18). Therefore, one solution is such that $b_{n^*} > 0$ for some n^* and $b_{n'} = 0$ for $n' \neq n^*$. Since $c'(q) > 0$ and $f_q(0;q) < 0$, $n^* \neq 0$. If $0 < n^* < N$, then by the FOC,

$$-b_{n^*} F_q(n^*-1;q) + b^* F_q(n^*;q) = c'(q). \tag{1.19}$$

Since $F_q(n^*-1;q) < 0$ and $F_q(n^*;q) < 0$, we can find an n^*-threshold contract t' and an (n^*+1)-threshold contract t such that $-t'_{n^*} F_q(n^*-1;q) = -t_{n^*+1} F_q(n^*;q) = c'(q)$ and $v(t;q) < v(t';q)$ by Lemma 4. By (1.19), we can find $\lambda > 1$ such that $b = \lambda t' + (1-\lambda)t$. Since $\lambda > 1$ and $v(t;q) < v(t';q)$, $v(t;q) < v(b;q)$, a contradiction. Therefore, it is necessary that $n^* = N$ and hence b is an N-threshold contract. Lastly, since IMCP holds, (b,q) is feasible in (1.2) and hence (b,q) solves (1.2).

If $q = 1$, then it is necessary that $F_q(N-1;1) < 0$. If otherwise $F_q(N-1;1) = 0$, then by DHRP, $F_q(n;1) = 0$ for all $0 \leq n < N$, but this implies no contract can incentivize $q = 1$, a contradiction. Since $F_q(N-1;q) < 0$, there exists an N-threshold contract that incentivizes $q = 1$. Since for $q < 1$, the cost-efficient contract is uniquely an N-threshold contract, the cost-efficiency of the N-threshold contract for $q = 1$ follows from continuity of $v(b;q)$ in both b and q. □

1.B.17 Proof of Proposition 15

Proof. If $N = 1$, then the statement is vacuously true since TSCP is not defined. Assume $N \geq 2$ and TSCP holds. The following lemma is useful in proving the proposition.

Lemma 5. *Let $f, g : [0,1] \to \mathbb{R}_+$ be two functions that are not constant and $f(0) = g(0) = 0$. Assume both f and g are continuously differentiable on $(0,1)$, with $g'(x) > 0$ for every $x \in (0,1)$. If $\frac{f'(x)}{g'(x)}$ is non-increasing for every $x \in (0,1)$, then $\frac{f(x)}{g(x)}$ is non-increasing on $(0,1)$.*

Proof of lemma. Since $g' > 0$ on $(0,1)$, $g(x) > 0$ for every $x \in (0,1)$.[18] With this observation and the assumption that f and g are both continuously differentiable, the condition that $\frac{f}{g}$ is

[18] I would like to thank Xiaohan Yan for the great discussion that leads to this proof.

non-increasing is equivalent to

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2} \leq 0.$$

Since $g > 0$ and $g' > 0$, the above inequality is equivalent to

$$s(x) = \frac{f'(x)}{g'(x)} - \frac{f(x)}{g(x)} \leq 0.$$

Therefore, $S = \{x \in [0,1] : s(x) > 0\}$ is the set on which $(f/g)' < 0$. To arrive at a contradiction, assume $S \neq \emptyset$. Since both f' and g' are continuous, S is open, and hence there exists some non-empty open interval $(a, b) \in S$. It remains to show that $a \in S$.

Since f'/g' is non-increasing by assumption and f/g is strictly increasing on (a, b), s is strictly decreasing on (a, b) and hence $\lim_{x \to a} s(x) > 0$. Since $a > 0$, $g'(a) > 0$, $s(a)$ is defined with $s(a) > 0$ and hence $a \in S$. If $\underline{s} = \inf S > 0$, then $\underline{s} \in S$ and there exists some $\epsilon > 0$ such that $\underline{s} - \epsilon \in S$ since S is open, contradicting the definition of $\underline{s}$. Therefore, $\inf S = 0$ and $(0, b) \in S$ for some $b > 0$. Since s is shown to be strictly decreasing on any open interval of S, $\lim_{x \to 0} s(x) > 0$.

Since $f(0) = g(0) = 0$, $f(x) = \int_0^x f'(y)dy$ and $g(x) = \int_0^x g'(y)dy$. For $x \in (0, 1)$, define $h(x) = f'(x)/g'(x)$. By assumption, h is non-increasing. In addition, since $f \geq 0$, $f(0) = 0$, and f is not constant, $f'(x) > 0$ on $(0, b)$ for some $b > 0$ and hence $h > 0$ on $(0, b)$. For every $x \in (0, b)$,

$$\frac{f(x)}{g(x)} = \frac{\int_0^x f(y)dy}{\int_0^x g(y)dy} = \frac{\int_0^x h(y)g(y)dy}{\int_0^x g(y)dy} \geq h(x)\frac{\int_0^x g(y)dy}{\int_0^x g(y)dy} = \frac{f'(x)}{g'(x)},$$

which implies $s(x) \leq 0$ on $(0, b)$ and there exists a sequence $(x_n)_n^\infty$ converging to 0 such that $s(x_n) \leq 0$ for every n. However, this is a contradiction to the earlier result that $\lim_{x \to 0} s(x) > 0$. Thus if f'/g' is non-increasing on $(0, 1)$, then so is f/g non-increasing on $(0, 1)$. The proof of the lemma is complete. □

Fix $0 < n < N$. Since TSCP holds, $-\frac{f_q(n;q)}{F_q(n;q)}$ is non-increasing in $q \in (0, 1)$. By A1(c), $f(n; 0) = 0$ and $1 - F(n; 0) = 0$. Additionally, $-F_q(n; q) > 0$ by A1(d); $f(n; q)$ and $F(n; q)$ are continuously differentiable by A1. The conditions of Lemma 5 are satisfied with $f(x) = f(n; x)$ and $g(x) = 1 - F(n; x)$, and hence $\frac{f(n;q)}{1-F(n;q)}$ is non-increasing in q on $(0, 1)$. Therefore, TSCP implies DHRP if A1 holds. □

Chapter 2

Flash Pass

[1]

2.1 Introduction

When access to a service facility is congested, service providers commonly implement a special type of queue called priority queue, where each person/entity in the queue has an associated priority such that those with a higher priority will be ahead of those with a lower priority in the queue. For example, an amusement park with a queue of customers can sell two types of priority passes to the customers, one called regular and the other called flash pass. A customer holding a flash pass is ahead of a customer holding a regular pass in the queue. There are many other applications of priority queue other than amusement parks. For example, for cloud computing, different computing requests queue for the computing resources; within a computer, different programs need to queue for access to the CPU; a popular museum has a queue of customers outside often let patrons with paid memberships skip the line; for large events hosting many visitors such as en exhibition, event organizers can let VIP pass holders cut in the line and enter the event venue ahead of regular pass holders. This paper considers the pricing problem faced by the seller that manages the priority queue, i.e., an amusement park.

When a park manages a priority queue by selling different priority passes, other customers' purchase decisions affect a customer's valuation of a pass, which makes the pricing problem of a priority queue different from one where a customer's valuation of a pass is fixed. For example, if many customer purchase a priority pass, then the congestion in that priority pass will lead to longer waiting time for customers with the same and lower priorities, lowering the valuation of passes with those priorities. Moreover, with the purchase decisions of other customers fixed, which priority pass a new customer will choose depends on the distribution of customers in each priority pass. In other words, the purchase behavior of one customer imposes externalities on other customers, and this paper aims to understand the implications of these externalities.

Specially, we observe that parks usually sell only a small number of priority passes. For example, Six Flags, a large amusement park corporation in the US, sells only three tiers of priority

[1]This is joint work with Yuichiro Kamada.

passes: THE FLASH Pass, THE FLASH Pass Gold, and THE FLASH Pass Platinum; Disneyland, a large theme park corporation with franchises all over the world, has two tiers of passes in terms of queuing priority: FASTPASS and regular ticket. Motivated by this observation, this paper focuses on the implementability of multi-pass schemes, i.e., whether a park can offer many priority passes and price them so that each pass has at least one willing customer. For example, this paper shows that when customers have the same utility function, the park cannot sell a different priority pass to each customer. We further show under some conditions on each customers' utility function, implementing a multi-pass scheme is not possible unless there exist large enough "gaps" between "adjacent" types. We formalize that the difficulty with selling many priority passes is due to the existence of externalities, which is explained in detail in the analysis.

This paper does not claim to thoroughly explain the constraint on the number of passes. The main objective of this paper is to show that externalities can contribute to the constrained number of priority passes, but other factors could contribute to the constraint in parallel. For example, selling many priority passes may incur considerable logistic costs, which could prompt the park to sell a small number of priority passes. Too many choices may also overwhelm customers (Park and Jang 2013; Kuksov and Villas-Boas 2010).

Queuing literature, such as Balachandran (1972), Adiri and Yechiali (1974), Hassin and Haviv (1997), and Alperstein (1988), has studied managing priority queues with setup different from this paper's. In these papers, each customer arrives sequentially, observes the state of the queues, makes a forecast about the waiting time for each pass, and then chooses a priority pass to maximize the customer's expected utility or leaves the queue. In particular, Alperstein (1988), who discusses the optimal pricing and the number of passes to sell, finds the profit-optimal number of passes equal to the number of customers, i.e., each customer is in a different priority pass. However, the number of priority passes is usually much smaller than the number of customers, and this paper provides one possible explanation for the disparity. In the queuing literature mentioned above, pricing many priority passes so that each priority has at least one customer does not pose a big difficulty: Adiri and Yechiali (1974) show that in equilibrium, early-arriving customers will purchase the lower-priority passes and later-arriving customers will purchase higher-priority passes as the queue gets large, which makes each priority pass have at least one customer.[2] This paper changes the assumption about each customer's information at decision-making time: each customer does not observe the state of the entire queue, whereas customers in the cited queuing studies do. We later show that this paper's setup, which is absent of the dynamic strategic consideration and changes each customer's payoff function, leads to a small number of priority passes.

This paper uses the mechanism-design approach to analyze the pass-selling problem, but it differs from the standard screening models in the following sense. In standard screening studies such as Guesnerie and Laffont (1984) and Maskin and Riley (1984) with a seller and buyers, the seller can manipulate the "quality" of a product, and a buyer's evaluation of a product does not depend on the purchase decisions of other buyers. This capability of the seller in the standard screening models, together with the setup that each buyer's valuation of a product is indepen-

[2] The intuition is that early arriving customers want to buy lower-priority passes because the queue is not congested at the time of arrival, whereas those arriving much later will want to buy the higher-priority passes in order to cut in the long queue.

dent of other buyers' decisions, allows the seller to separate different types of customers. In contrast, in this paper, the park cannot fully control the "quality" of a pass, and how much a customer values a pass depends on other customers' purchase decisions.[3] In short, the externality among customers, which is characteristic of queue management, makes our model different from the standard screening models. Because the standard screening studies just mentioned do not consider the externalities among customers, which contributes to the park's inability to have full control over the "qualities" of the priority passes, implementability in this paper cannot be extended from these studies: the externality makes implementation harder, which Section 2.6.2 discusses in detail.

Several existing studies consider externalities. Segal (1999) considers the bilateral contracting between a principal and a number of agents. To each agent, the principal offers a contract that specifies the trade quantity between the principal and the agent. The agent then chooses whether to accept the contract.[4] The contracting outcomes of other agents can affect the reserve utility of an agent, and that paper provides sufficient conditions about the externalities under which the principal-optimal aggregate trade quantity would be above or below the socially optimal aggregate trade quantity.[5] If we wish to fit our model to the setup of Segal (1999), given a priority queue, we can treat a customer's priority pass as the trade profile[6] and the optimal deviation payoff of the customer as a customer's reserve utility. The externalities in the fitted model do not satisfy the conditions for the results of Segal (1999), and hence that paper's results do not cover ours.

For existing applied literature that includes externalities, the nature of externalities varies by application. For example, Katz and Shapiro (1986) analyze the adoption of new technology under the presence of network effects that greater adoption of the technology increases the utility of adoption; Csorba (2008) considers the externalities when the utility of using a product increases with the rise in demand; Shi, Zhang, and K. Srinivasan (2019) and Kamada and Öry (2020) both consider product-line design and pricing questions in which new customers boost the utility of existing customers. These papers and this paper differ by the form of externalities: each paper

[3]Even though to some extent the park can fine-tune the prices to adjust the quality, the park is still not in full control. For example, lowering the price of a higher-priority pass attracts more customers from the lower-priority passes, creating more congestion in the higher-priority pass and hence lowering the quality of that pass. However, at the same time, the quality of a lower-priority pass is affected as well, because to customers staying in the lower-priority pass, there are now even more customers with a strictly higher priority, hence lowering the quality of the lower-priority pass.

[4]Segal and Whinston (2003), in a follow-up paper, considers the convergence to competitive equilibrium in vertical contracting when the principal and each agent contract on a private menu. With respect to multilateral contracting, Gomes (2005) and Bloch and Gomes (2006) consider dynamic multilateral bargaining among agents forming coalitions and bargaining to split coalition-structure-dependent surplus.

[5]Specifically, Segal (1999) finds that when the externalities are positive (negative), i.e., an agent's reserve utility is non-decreasing (non-increasing) if other agents are trading more with the principal, then the principal-optimal aggregate trade quantity will be below (above) the socially optimal aggregate trade quantity. See Bergstrom, Blume, and Varian (1986) for an application with positive externalities and Rasmusen, Ramseyer, and Wiley Jr (1991) for an application with negative externalities.

[6]A trade profile is a collection that can be mapped to an outcome. For example, under this paper's context, a trade profile can be considered as the collection of each customer's priority pass. This collection of priority passes is then mapped to the resulting priority queue as the outcome.

models externalities to cater to that paper's application, and none of the papers cited in this paragraph deals with externalities in queues. In this paper, the externality that plays a crucial part in results of this paper is the externality created when a customer switches to a different priority pass with other customers' purchase decisions fixed.

Section 2.2 and 2.3 introduce the main model. Section 2.4 discusses selling multiple priority passes with one utility type. Section 2.5 extends the main model to cases with multiple types of utility functions. Section 2.6 provides a more detailed discussion of the derived results, with Section 2.7 concluding the paper.

2.2 Model

An amusement park sells $K \geq 1$ different types of priority passes, $\{\theta_k\}_{k=1}^K$, to $N \geq 1$ customers, where θ_k is the k-th highest priority pass. Let θ_0 denote the option of leaving the queue. The park sets $p = (p_1, \dots, p_K) \in \mathbb{R}_+^K$ where p_k is the price of θ_k. For completeness, set $p_0 = 0$. Having observed p, customers make purchase decisions simultaneously: each customer either buys some priority pass or leaves the park with reserve utility denoted by u_0, which is set to zero unless otherwise specified.[7]

After the purchase decisions, the customers that purchase a priority pass form a queue, with the possible positions in the queue being $1, 2, \dots, N$. A customer's valuation of visiting an amusement park depends solely on the customer's position in the queue and the price the customer pays for the priority pass. A *base utility function* $u : \mathbb{N} \to \mathbb{R}$ assigns a utility to each position in the queue, where u_n denotes the utility from being at the n-th position in the queue.[8]

For customer i, let $A_i = \{\theta_k\}_0^K$ denote the set of customer's possible purchase decisions and define $A = \times_{i=1}^N A_i$. Given $a \in A$, define $\bar{q}(a) = (\bar{q}_k(a))_{k=0}^K$, where $\bar{q}_k(a) = |\{i : a_i = \theta_k\}|$, which is the number of customers choosing θ_k. If $1 \leq k \leq K$ and $a_i = \theta_k$, the customer is guaranteed to be ahead of every customer in θ_j if $j > k$ and behind every customer in θ_j if $1 \leq j < k$.[9] For customers in the same priority pass, assume each order of these customer happens with the same probability. See Figure 2.1 for an example of priority queue. This assumption implies that to each customer in a priority pass, the customer's position in the queue is uniformly distributed on the possible positions of that priority pass.

Let $\mu_i : A \to \mathbb{R}$ denote the utility of customer i from a strategy profile. Given $q = \overline{q}(A)$, if $1 \leq k \leq K$, let $Q_k(q) = \sum_{j=1}^k q_j$ denote the last position of θ_k and set $Q_0(q) = 0$. The assumptions

[7]In Section 2.6.3, we let the number of customers grow toward infinity, and in this case, the reserve utility is non-zero.

[8]To clarify, $\mathbb{N}$ denotes the set of strictly positive integers.

[9]Such priority queue is called preemptive as every higher-priority customer is ensured a faster entry than a lower-priority customer. Not all priority queues are preemptive and whether preemptive priority queue is optimal does not fall under the scope of this paper. While this paper's setup does not fit perfectly to the applications in this paper, it can capture some important aspects of implementing a priority queue.

Customer	Pass Bought	Price Paid	Expected Utility
A	θ_1	p_1	$\frac{u_1+u_2}{2} - p_1$
B	θ_1	p_1	$\frac{u_1+u_2}{2} - p_1$
C	θ_2	p_2	$\frac{u_3+u_4+u_5}{3} - p_2$
D	θ_2	p_2	$\frac{u_3+u_4+u_5}{3} - p_2$
E	θ_2	p_2	$\frac{u_3+u_4+u_5}{3} - p_2$
F	θ_3	p_3	$u_6 - p_3$
G	θ_0	$p_0(=0)$	0

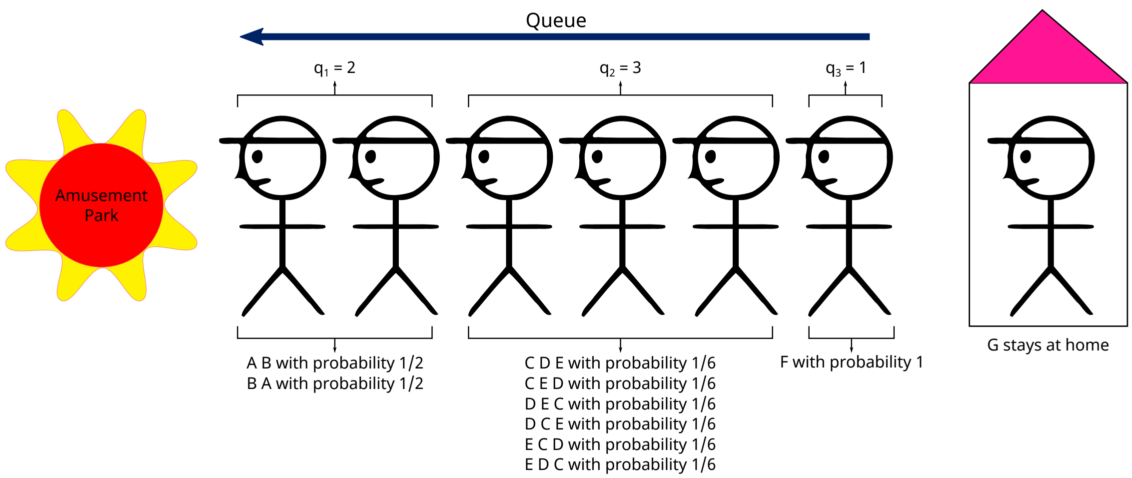

Figure 2.1: Example of a scheme: $(q_0, q_1, q_2, q_3) = (1, 2, 3, 1)$.

made so far imply that μ_i can be written as

$$\mu_i(a) = \begin{cases} \dfrac{\sum_{n=Q_{k-1}(\bar{q}(a))+1}^{Q_k(\bar{q}(a))} u_n}{\bar{q}_k(a)} & \text{if } a_i = \theta_k \neq \theta_0 \\ u_0 & \text{if } a_i = \theta_0 \end{cases}.$$

Note that each customer position in the queue and the customer's utility depend on other customers' purchase decisions. Given N and K, define $\mathcal{Q}(N, K) = \{q \in (0 \cup \mathbb{N}) \times \mathbb{N}^K \,:\, \sum_{k=0}^{K} q_k = N\}$ to be the set of *schemes* for (N, K). If $q_0 > 0$, then some customers stay at home and do not join the queue. The assumption that $q_k > 0$ if $k > 1$ means that the model requires a scheme to have at least one customer for each priority pass. See Figure 2.1 for an example that illustrates the setup of the model.

The above setup can be modeled as a strategic-form game. Given the tuple (N, K, p, u), define a strategic-form game $G(N, K, p, u) = \langle N, A, (\pi_i)_{i=1}^N \rangle$, where $\pi_i : A \to \mathbb{R}$ is each customer's (common) payoff function, which is assumed to be of a quasi-linear form: for every $a \in A$, $\pi_i(a) = \mu_i(a) - p(a_i)$, where $p(a_i)$ is the price paid by customer i, i.e., $p(a_i) = p_k$ if $a_i = \theta_k$.

2.3 Preliminaries

We now define *implementability*, the main concept of this paper. In short, a scheme is implementable if it results from each customer's optimal purchase decision that factors in other customers' decisions.

Definition 4 (Implementation)**.** Fix (N, K, u). A price vector p *implements* a scheme $q \in \mathcal{Q}(N, K)$ if $G(N, K, p, u)$ has a (pure-strategy) Nash equilibrium a^* such that $\bar{q}(a^*) = q$. A scheme q is *implementable* for (N, K, u) if there exists a price vector p that implements q.

When (N, K, u) is without ambiguity, we only write "q is implementable" without including (N, K, u). To give an intuitive understanding, an equivalent formulation of implementation based on incentive constraints is provided. We first define the utility of deviating from a strategy profile. Given $q \in \mathcal{Q}(N, K)$, define $\overline{v}(\cdot; q) : \{\theta_k\}_{k=0}^K \to \mathbb{R}$ such that $\overline{v}(\theta_0; q) = u_0$ and $\overline{v}(\theta_k; q) = \sum_{n=Q_{k-1}+1}^{Q_k} u_n / q_k$ if $k \neq 0$, which is the average utility of positions of θ_k in q.

Definition 5 (Pass-Utility function)**.** Fix (N, K, u) and a scheme $q \in \mathcal{Q}(N, K)$. For each j and k from $\{0, \dots, K\}$ such that $q_j > 0$, define $v(\theta_k; \theta_j; q) = \overline{v}(\theta_k; q')$, where q' is the resulting scheme when one customer holding θ_j in q switches to θ_k. We call v the *pass-utility function constructed from u.*

In words, $v(\theta_k; \theta_j; q)$ gives the utility a customer will get (before payment) if the customer switches from θ_j to θ_k. For example, given a scheme q, if $1 \leq j < k \leq K$ and a customer switches from θ_j to θ_k, then the positions of θ_k in the new queue are from Q_{k-1} to Q_k, and thus $v(\theta_k; \theta_j; q) = \sum_{n=Q_{k-1}}^{Q_k} u_n / (1 + q_k)$.

When without ambiguity, such as when the scheme in consideration is fixed, q is omitted and $v(\theta_k; \theta_j)$ is written instead. Abuse notation to write $v(\theta) := v(\theta; \theta)$ for each $\theta \in \{\theta_k\}_{k=0}^K$. In words, $v(\theta)$ denotes the utility of a customer choosing θ in a scheme. The following result illustrates some important properties of the pass utility function, which is fundamental to the results in this model.

Claim 1 (Properties of pass-utility function)**.** *Fix* (N, K, u) *and* $q \in \mathcal{Q}(N, K)$. *If* $1 \leq j_1 \leq j_2 < k \leq l_1 \leq l_2 \leq K$*, then*

$$v(\theta_k; \theta_{j_1}) = v(\theta_k; \theta_{j_2}) > v(\theta_k) > v(\theta_k; \theta_{l_1}) = v(\theta_k; \theta_{l_2}). \tag{2.1}$$

The first equality means that the utility of an downgrade does not depend on "how much" higher priority the higher-priority pass has than the lower-priority pass; similarly, the second equality means the utility of an upgrade does not depend on "how much" higher priority the higher-priority pass has than the lower-priority pass. Additionally, the two inequalities mean that a downgrade improves the utility of the lower-priority pass and an upgrade lowers the utility of the higher-priority pass. The intuition is that when a customer downgrades to a lower-priority pass, say θ_k, the first position is θ_k improves by one and the last position of θ_k is unchanged, leading to a higher average utility of positions of θ_k; when a customer upgrades to a higher-priority pass θ_k, the first position of θ_k is unchanged and the last position of θ_k decreases by one, lowering the average utility of positions of θ_k.

We now define implementation with respect to customers' incentive constraints. Fix (N, K, u) and i such that $1 \leq i \leq N$. Fix $q \in \mathcal{Q}(N, K)$ and a price vector p. For j from $\{1, \dots, K\}$, (p, q) is said to satisfy the *individual-rationality constraint of* θ_j (henceforth IR_j) if every customer buying θ_j in q has no incentive to leave the queue, i.e.,

$$v(\theta_j) - p_j \geq u_0 = 0. \tag{IR$_j$}$$

Let the *set of IR constraints* be the collection of IR_j over j such that $1 \leq j \leq K$. Secondly, for j and k such that $0 \leq j \leq K$ and $1 \leq k \leq K$, (p, q) is said to satisfy the *incentive-compatibility constraint from* θ_j *to* θ_k (henceforth IC_{jk}) if every customer with θ_j in q has no incentive to switch to θ_k, i.e.,

$$v(\theta_j) - p_j \geq v(\theta_k; \theta_j) - p_k. \tag{IC$_{jk}$}$$

Let the *set of IC constraints* be the collection of IC_{jk} over j and k such that $0 \leq j \leq K$, $1 \leq k \leq K$, and $q_j > 0$.[10] We now show that implementability can be defined with respect to IC and IR constraints.

Claim 2 (Implementation with respect to incentive constraints). *A scheme $q \in \mathcal{Q}(N, K)$ is implementable if and only if there exists a price vector p such that (p, q) satisfies every constraint in the set of IC and IR constraints.*

Fix an IC constraint IC_{jk} where $1 \leq j < k \leq K$. The constraint IC_{jk} is a *downward IC constraint* if $j < k$, a *local downward IC constraint* if $k = j + 1$, an *upward IC constraint* if $j > k$, and a *local upward IC constraint* if $j = k + 1$. In standard screening models such as Guesnerie and Laffont (1984) and Maskin and Riley (1984), the set of incentive constraints can be reduced to IR_K and the set of local downward IC constraints.[11] Similarly, we will show that the set of downward IC constraints can be reduced to the set of local downward IC constraints: IC_{jk} holds if $\text{IC}_{l,l+1}$ holds for every l such that $j \leq l < k$. Unlike the standard screening models, however, the set of upward IC constraints cannot be simplified to the set of local upward IC constraints. The following lemma summarizes the constraint reduction results in this model and provides a condition that is useful for proving some negative results about implementation.

Lemma 6 (Constraint reduction). *Fix $q \in \mathcal{Q}(N, K)$ and a price vector $p \in \mathbb{R}_+^K$. Fix j and k such that $1 \leq j < k \leq K$.*

[10] If $q_j = 0$, then IC_{jk} does not need to hold for q to be implementable. By definition, $q_j > 0$ if $j > 0$.

[11] Carroll (2012) provides a comprehensive discussion about when local incentive constraints are sufficient for global incentive constraints. This paper's setup is most closely related to the case with transfers and interdependent preferences in that paper. Carroll (2012) shows that in that case, if each agent's utility is linear in the reported type and each agent's type space is convex, then local incentives are sufficient. The conditions cover screening studies like Guesnerie and Laffont (1984) and Maskin and Riley (1984), but not this paper because the linearity condition does not hold in our model.

(a) If (p, q) satisfies $\text{IC}_{l,l+1}$ for every l such that $j \leq l < k$, then (p, q) satisfies IC_{jk}.

(b) If (p, q) satisfies IR_k and IC_{jk}, then (p, q) satisfies IR_j.

(c) If (p, q) satisfies both IC_{jk} and IC_{kj}, then

$$v(\theta_j; \theta_k) - v(\theta_k) \leq v(\theta_j) - v(\theta_k; \theta_j). \tag{ID$_{jk}$}$$

A scheme q is said to satisfy ID_{jk} if the inequality above holds. Let the set of *increasing difference (ID)* conditions be the collection of ID_{jk} for j and k such that $1 \leq j < k \leq K$. Note that ID_{jk} holds only if IC_{jk} and IC_{kj} hold simultaneously. Thus the set of ID conditions is necessary for implementation. To see the intuition, note that IC_{jk}, under which every customer holding θ_j has no incentive to downgrade to θ_k, implies an upper bound on $p_j - p_k$, whereas IC_{kj}, under which every customer holding θ_k has no incentive to upgrade to θ_j, implies a lower bound on $p_j - p_k$. Implementation necessitates that the upper bound must be weakly higher than the lower bound, and hence ID_{jk}. The following simple example illustrates this intuition.

Example 5 (ID condition)**.** Consider a case with $N = 3$, $K = 2$, $(q_0, q_1, q_2) = (0, 1, 2)$. Assume there exists a price vector (p_1, p_2) that implements q. Note that IC_{12}, under which a customer holding θ_1 has no incentive to switch to θ_2, implies

$$p_1 - p_2 \leq v(\theta_1) - v(\theta_2; \theta_1) = u_1 - \frac{u_1 + u_2 + u_3}{3} = \frac{2u_1 - u_2 - u_3}{3},$$

while IC_{21}, under which a customer holding θ_2 has no incentive to switch to θ_1, implies

$$p_1 - p_2 \geq v(\theta_1; \theta_2) - v(\theta_2) = \frac{u_1 + u_2}{2} - \frac{u_2 + u_3}{2} = \frac{u_1 - u_3}{2}.$$

The two inequalities together imply

$$\frac{u_1 - u_3}{2} = v(\theta_1; \theta_2) - v(\theta_2) \leq p_1 - p_2 \leq v(\theta_1) - v(\theta_2; \theta_1) = \frac{2u_1 - u_2 - u_3}{3}, \tag{2.2}$$

which is equivalent to ID_{12}.

The ID conditions are commonly assumed in the screening literature. In existing studies such as Maskin and Riley (1984) and Guesnerie and Laffont (1984), the ID conditions are necessary and sufficient for implementation if the set of IC constraints is reduced to the set of local IC constraints.[12] In this paper, while the ID conditions are necessary for implementation, they are not sufficient because constraint reduction does not work for the set of upward IC constraints. The only special case where the ID conditions are necessary and sufficient for implementation is when $K = 2$ with every customer buying some pass because $K = 2$ is the only case where the set of local IC constraints is the same as the set of IC constraints. For $K \geq 3$, the ID conditions are not sufficient in this model, and thus conditions more than just the ID conditions are needed, which the next section discusses.

[12] See Chapter 2 of Bolton and Dewatripont (2005) for a summary of these results.

2.4 Implementability

This section discusses the implementability of schemes with different patterns of base utility functions. Although some results, such as Proposition 20, are applicable to more general base utility functions, we pay special attention to the following three patterns of base utility functions, each of which has reasonable applications and hence is worth some analysis.

- Concave base utility function ($u_n - u_{n+1} \leq u_{n+1} - u_{n+2}$ for each n): the concave case applies when queueing has an opportunity cost to a customer. Specifically, assume that a customer's utility from a park depends on the time spent in the park. Assume the customer has a total amount of available time T. If the customer spent x units of time in the park, where $0 \leq x \leq T$, then the customer's utility would be $y(x)$, where $y' > 0$ and $y'' < 0$, i.e., the customer enjoys spending time in the park but faces diminishing marginal utility. Assuming that being at n-th position means that the customer will wait for n units of time before going into the park, the customer's utility is $u_n = y(T - n)$, which is decreasing and concave in n.

- Linear base utility function ($u_n - u_{n+1} = u_{n+1} - u_{n+2}$ for each n): the linear case is commonly assumed in queuing literature from operations research, e.g., Balachandran (1972), Adiri and Yechiali (1974), and Alperstein (1988).

- Convex base utility function ($u_n - u_{n+1} \geq u_{n+1} - u_{n+2}$ for each n): the convexity assumption is applicable when queuing is inherently unpleasant and a customer becomes less sensitive to longer queuing time. Additionally, the convexity assumption holds when upon entry, the customer obtains a fixed instantaneous utility that is exponentially discounted with respect to queuing time that is linear in the positions.

Start with the concave base utility functions. It turns out when the base utility function is strictly concave, i.e., $u_n - u_{n+1} < u_{n+1} - u_{n+2}$ for every position n, then no scheme with more than one pass and more than two customers is implementable.

Proposition 18 (Implementation with strictly concave utility)**.** *Fix (N, K, u) where $K > 1$, $N > 2$, and u is strictly concave. If $q \in \mathcal{Q}(N, K)$, then q is not implementable.*

The reason for this negative result for strictly concave base utility functions is that every scheme with more than one pass necessarily violates some ID condition. The following example illustrates the intuition.

Example 6 (Example 5 continued)**.** Assume u is strictly concave, i.e., $u_n - u_{n+1} < u_{n+1} - u_{n+2}$ for $1 \leq n \leq N$. For the scheme q to be implementable, ID_{12} in (2.2) requires $\frac{u_1 - u_3}{2} \leq \frac{2u_1 - u_2 - u_3}{3}$, which implies $u_1 - u_2 \geq u_2 - u_3$, which contradicts the strict concavity assumption. Hence ID_{12} is violated.

The key intuition is that each customer has some positive size. In other words, when a customer switches to a different pass, externalities are created in the new pass. Indeed, in the hypothetical situation without externalities, i.e., $v(\theta_k; \theta_j) = v(\theta_k)$ for every j and k, IC_{12} for the aforementioned example implies $p_1 - p_2 \leq u_1 - \frac{u_2+u_3}{2}$ and IC_{21} implies $p_1 - p_2 \geq u_1 - \frac{u_2+u_3}{2}$, making the scheme implementable as in the standard screening models.

With a simple modification of the proof of Proposition 18, it can be shown that when u is convex, i.e., $u_n - u_{n+1} \geq u_{n+1} - u_{n+2}$ for every position n, all ID conditions are satisfied. Since the ID conditions can be shown to be sufficient for implementation with $K = 2$, for convex u, every scheme is implementable, as follows.

Proposition 19 (Two-pass implementation with convex utility). *Fix (N, K, u) where $K = 2$ and u is a convex base utility function. Fix $q \in \mathcal{Q}(N, 2)$. If $v(\theta_2) \geq u_0$, then q is implementable.*

When $K \geq 3$, not every scheme is implementable. For example, we show later in Proposition 21 that when the base utility function is linear, which is a special case of convexity, no schemes with three or more priority passes are implementable. The different implementability results between Proposition 18 and Proposition 19 show that implementability depends on the shape of u. It turns out that to check implementability of a scheme, it is sufficient and necessary to check whether the price vector binding all local downward IC constraints implements the scheme.

Lemma 7 (Implementation condition). *Fix (N, K, u) where $1 \leq K \leq N$. Let $q \in \mathcal{Q}(N, K)$ be such that $q_0 = 0$. Let p be a price vector such that $p^*_K = v(\theta_K)$ and $p^*_k - p^*_{k+1} = v(\theta_k) - v(\theta_{k+1}; \theta_k)$ for every k such that $1 \leq k < K$. The scheme q is implementable if and only if (p^*, q) satisfies every upward IC constraint and $v(\theta_K) \geq 0$.*

The lemma above is relatable to Theorem 1 in Rochet (1987). In that paper, the necessary and sufficient condition for a scheme q to be implementable is that given any finite cycle of passes, $k_0, k_1, \dots, k_m, k_{m+1} = k_0$ in $\{0, \dots, K\}$, if $\text{IC}_{k_l k_{l+1}}$ is bound for $l = 0, \dots, m-1$ with the implied price difference $p_{k_0} - p_{k_m}$, then this price difference must satisfy $\text{IC}_{k_m k_{m+1}}$. Lemma 6 shows that if IR_K and every local downward IC constraint are bound, then all downward IC and IR constraints hold. Hence in the above lemma, it only remains to check whether given the scheme q, (p^*, q) satisfies every upward IC constraint.

In the standard screening models, with a fixed scheme q, p^* binds every local downward IC constraint, and then every IC and IR constraint holds thanks to the ID conditions and constraint reductions. What differentiates this paper's model is that the set of upward IC constraints cannot be reduced to the set of local upward IC constraints, hence the necessity of Lemma 7.

Lemma 7 also illustrates how a scheme fails to be implementable. If $1 \leq j < k \leq K$,

$$p^*_j - p^*_k = \sum_{l=j}^{k-1} p^*_l - p^*_{l+1} = v(\theta_j) - v(\theta_k) - \sum_{l=j}^{k+1} v(\theta_{l+1}; \theta_l) - v(\theta_{l+1}), \tag{2.3}$$

where p^* is defined in Lemma 7, and the summation is the accumulated externalities from downgrading to the next-lower-priority pass. At the same time, IC_{kj} implies

$$p_j - p_k \geq v(\theta_j; \theta_k) - v(\theta_k) = v(\theta_j) - v(\theta_k) - \left[v(\theta_j) - v(\theta_j; \theta_{j+1}) \right], \tag{2.4}$$

where the second difference term is the externality of upgrading from θ_{j+1} to θ_j. Whenever the sum of downgrade externalities in (2.3) exceeds the upgrade externality in (2.4), the scheme is not implementable.

In Alperstein (1988), it profit-optimal to set the number of priority passes equal to the number of customers. The following result shows that this scheme is not implementable in our model.

Proposition 20 (Implementation with $N = K$)**.** *Fix* (N, K, u) *and* $q \in \mathcal{Q}(N, K)$*. If* $K > 2$ *and* $q_k = 1$ *for every* k *such that* $1 \leq k \leq K$*, then* q *is not implementable.*

The result is shown by contradiction. Towards this end, suppose that there is such a scheme that is implementable and consider the incentives of the first three customers, each of whom buys a different priority pass. For the first customer, it is tempting to switch to θ_1 from θ_2, since the customer pays less but still has the chance of being at the same position after switching. To incentivize the first customer from downgrading, $p_1 - p_2$ needs to be small. Similarly, $p_2 - p_3$ needs to be small so that the second customer does not want to downgrade to θ_3 from θ_2. For the third customer, however, upgrading to the first pass can be tempting since by doing so, the third customer strictly improves the customer's position in the queue, albeit at a higher price. Hence $p_1 - p_3$ needs to be large enough to disincentivize the customer from upgrading to θ_1 from θ_3. The proof shows that the upper bound on $p_1 - p_3$ implied by the upper bounds on $p_1 - p_2$ and $p_2 - p_3$ is strictly less than the lower bound on $p_1 - p_3$, which is a contradiction.

Another special case of the base utility function u is the linear utility function. Since the linear case is weakly convex, by Proposition 19, every scheme with $K = 2$ is implementable. For $K > 2$, it can be shown that no scheme is implementable.

Proposition 21 (Implementation with linear utility)**.** *Fix* (N, K, u)*, where* u *is linear, i.e.,* $u_n - u_{n+1} = d > 0$ *for every* n *such that* $1 \leq n < N$*. Fix* $q \in \mathcal{Q}(N, K)$ *such that* $v(\theta_K) \geq u_0$*. The scheme* q *is implementable if and only if* $K \leq 2$*.*

The proof shows that when $K \geq 3$, IC_{31} implies $p_1 - p_3 \geq v(\theta_1) - v(\theta_3) - \frac{d}{2}$, but binding IC_{12} and IC_{23} implies $p_1 - p_3 = v(\theta_1) - v(\theta_3) - d$, arriving at a contradiction. Therefore, for linear base utility functions, when the local downward IC constraints are bound, the externalities created by downgrades lower p_1 so much that the lowest-priority pass customers have an incentive to upgrade.

We have assumed that all customers have the same base utility function. The next section considers the case where customers have different base utility functions. We are going to show that although the heterogeneity in the utility functions makes it possible to implement a scheme with more than two passes as Six Flag does, the conflict between the upgrade and downgrade incentives can persist when there are multiple types of base utility functions.

2.5 Extension to Heterogeneous Utilities

Name the case with one base utility function the *single-type case.* Now consider the case where customers have heterogeneous base utility functions, which we call the *multi-type case.* To be precise, assume each customer's base utility function comes from $\{u^t\}_{t=1}^T$, where $1 \leq t \leq T$ and t is the index for utility type. For each $t = 1, \dots, T$, let N^t be the number of customers with base utility function u^t. Let $N = \sum_{t=1}^{T} N^T$. Assume that the reserve utility of each customer type is zero, i.e., $u_0^t = 0$ for all type t.[13] Let $G((N^t)_{t=1}^T, K, p, (u^t)_{t=1}^T)$ be the strategic-form game defined analogously to that in the single-type case. Given (N, K), let $\overline{\mathcal{Q}}(N, K) = \{q \in (\{0\} \cup \mathbb{N})^{K+1} : \sum_{k=1}^{K} q_k = N\}$. Define the set of schemes as

$$\mathcal{Q}\left(\left(N^t\right)_{t=1}^{T}, K\right) = \left\{\left(q^t \in \overline{\mathcal{Q}}(N^t, K)\right)_{t=1}^{T} : \sum_{\tau=1}^{T} q_k^\tau > 0 \text{ if } 1 \le k \le K\right\}.$$

The restriction that $\sum_{t=1}^{T} q_k^t > 0$ if $1 \le k \le K$ ensures that every priority pass has at least one customer, which is analogous to the definition in the single-type case. For each customer type t, the pass utility function v^t is constructed from u^t. Given j and k such that $0 \le j \le K$ and $1 \le k \le K$, IC_{jk}^t and IR_k^t are defined analogously to the single-type case with respect to v^t. Different from the single-type case, IC_{jk}^t and IR_k^t may not need to hold for implementation. To be precise, in an implementable scheme, IC_{jk}^t needs to hold only if there is some customer of type t in θ_j, i.e., $q_j^t > 0$. Similarly, IR_k^t needs to holds only if $q_k^t > 0$.

Given $\{u^t\}_{t=1}^T$, assume that if $1 \le \tau_1 < \tau_2 \le T$, then at every position n, $u_n^{\tau_1} > u_n^{\tau_2}$ and $u_n^t - u_{n+1}^t > u_n^{\tau_2} - u_{n+1}^{\tau_2}$. A customer of type τ_1 is said to be of a (weakly) higher type than a customer of type τ_2 if $\tau_1 \le \tau_2$.[14]

2.5.1 Two Utility Types

For now, consider the case where there are two types of base utility functions called the high and low types. Denote the high type's base utility function by u^h and that of the low type's by u^l. Fix $\left((N^h, N^l), K\right)$ and a scheme $q \in \mathcal{Q}((N^h, N^l), K)$. For j and k such that $0 \le j \le K$ and $1 \le k \le K$, both IC_{jk}^h and IC_{jk}^l are constraints about upper bounds on $p_j - p_k$, and hence there must exist one constraint that implies the other. Specifically, by the construction of types, if $1 \le j < k \le K$, then IC_{jk}^l implies IC_{jk}^h; if $1 \le k < j \le K$ or $j = 0$, then IC_{jk}^h implies IC_{jk}^l.

With the above observation that one IC_{jk}^t for some type t is sufficient to describe all the incentives of switching from θ_j to θ_k, we can define IC_{jk} as we did in the single-type case. Specifically, fix j and k such that $0 \le j < k \le K$. If $j > 0$ and $q_j^l > 0$, let $\text{IC}_{jk} = \text{IC}_{jk}^l$; if $j = 0$ and $q_j^h > 0$, let $\text{IC}_{jk} = \text{IC}_{jk}^h$; if $j > 0$ and $q_j^l = 0$, let $\text{IC}_{jk} = \text{IC}_{jk}^h$. If $j > 0$ and $q_k^h > 0$, let $\text{IC}_{kj} = \text{IC}_{kj}^h$; if $j > 0$ and $q_k^h = 0$, let $\text{IC}_{kj} = \text{IC}_{kj}^l$. Similarly, we can define IR_k. Specifically, if $q_k^l > 0$, then let $\text{IR}_k = \text{IR}_k^l$; if $q_k^l = 0$, let $\text{IR}_k = \text{IR}_k^h$.

[13] The homogeneous reserve utility assumption introduces some loss of generality as constraint reduction results of this section need to additionally take care of each type's different reserve utility. Since introducing heterogeneous reserve utilities further complicates the analysis without providing new insights about the pass-switching incentives, this paper makes the simplifying homogeneous reserve utility assumption.

[14] A customer type τ_1 is said to be higher than τ_2 in the sense that the utility of customers of type τ_1 decreases faster than those of type τ_2 with respect to positions in a queue. We assign lower indices to higher customer types because results introduced later show that in implementable schemes, customers of higher customer types necessarily hold higher-priority passes and so they occupy lower-indexed positions in a queue.

Let the set of IC constraints be the collection of IC_{jk} over j and k such that $0 \leq j \leq K$, $1 \leq k \leq K$ and $q_j^h + q_j^l > 0$; let the set of IR constraints be the collection of IR_k over k such that $1 \leq k \leq K$. With this notation, implementation is defined with respect to the set of IR and IC constraints just like the single-type case. For $1 \leq j < k \leq K$, let $\underline{t}_j$ be the customer type such that $\text{IC}_{jk}^{\underline{t}_j} = \text{IC}_{jk}$ and $\bar{t}_k$ be the customer type such that $\text{IC}_{kj}^{\bar{t}_k} = \text{IC}_{kj}$.[15]

The addition of this heterogeneity introduces some complications for constraint reduction. In the single-type case, Lemma 6 and 7 reduce the set of downward IC constraints to the set of local downward IC constraints and the set of IR constraints to only IR_K. Generally, neither types of constraint reductions holds in the two-type case, but there are conditions under which the reductions of IR and downward IC constraints hold.[16] It turns out that the constraint reduction results as in the single-type case can be obtained if we impose the restriction that $q_j^l > 0$ implies $q_k^l > 0$ if $1 \leq j \leq k \leq K$. This restriction eliminates schemes where a low-priority pass has only high-type customers but a high-priority pass has at least one low-type customer.

Now focus on the linear utility, a special case of concave utility functions. If a scheme q is implementable, then $K \leq 4$ by the proof of Proposition 21.[17] The next result presents the conditions under which there exists $q \in \left((N^h, N^l), K, (u^h, u^l)\right)$ that is implementable for $K = 4$. The intuition of the second condition in the result is that the slopes of u^h and u^l need to be sufficiently different from each other.

Theorem 5 (Implementation with Two Linear Utilities). *Consider $\left((N^h, N^l), K, (u^h, u^l)\right)$, where $K = 4$. Suppose $u_n^l = \alpha^l - nd$ and $u_n^h = \alpha^h - \beta^h nd$, where $d > 0$, $u_{N^h+N^l}^h - u_{N^h+N^l}^l > 0$, and $\beta^h > 1$. Consider $q = ((q_1^h, q_2^h), (q_1^l, q_2^l)) \in \mathcal{Q}\left((N^h, N^l), K\right)$ such that $q_0^h = q_0^l = 0$. Assume $v^l(\theta_K) \geq 0$. There exists $b \geq 0$ such that q is implementable if and only if the following two conditions hold:*

(a) $q_1^h + q_2^h = N^h$ *and* $q_1^l + q_2^l = N^l$.

(b) $\beta^h \geq b$.

Part (a) of Theorem 5 is a special case for the restriction that $1 \leq j < k \leq K$, then $q_j^l > 0$ implies $q_k^l > 0$. The restriction is used for the constraint reduction results. As implied by Theorem 5, it turns out that the restriction is necessary for implementation with two concave utilities. To be precise, implementation with two concave base utility functions implies a condition stronger than the restriction for constraint reduction.

Proposition 22 (Implementable schemes with two concave utilities). *Fix $\left((N^h, N^l), K, (u^h, u^l)\right)$ such that both u^h and u^l are concave. If $q \in \left((N^h, N^l), K\right)$ is implementable and $1 \leq j < k \leq K$, then $q_j^l > 0$ implies $q_k^h = 0$.*

[15]The choice of type is unique as $u_n^h - u_{n+1}^h > u_n^l - u_{n+1}^l$ at every position n.

[16]See Lemma 8 and Lemma 9 in Appendix 2.A.

[17]If $K > 4$, there is at least one customer type with customers in three different passes. The same argument as in Proposition 21 shows that this is not implementable.

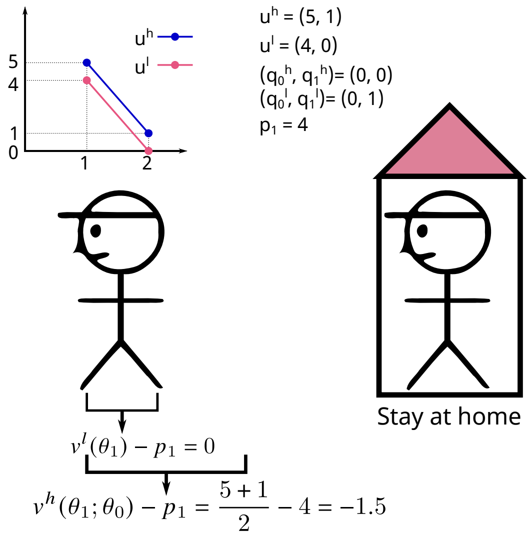

Figure 2.2: An example where low type buys some pass but high type does not.

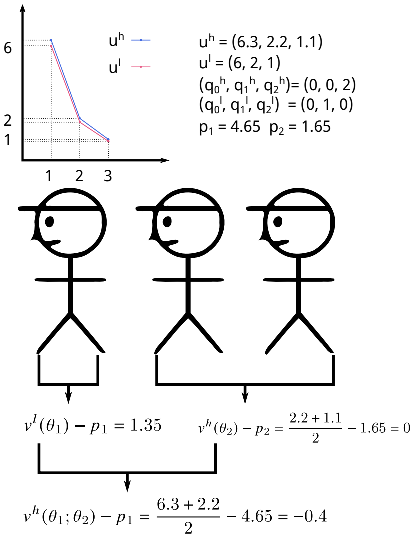

Figure 2.3: An example where high type buys some strictly lower-priority pass than low type.

The above condition is relatable to a monotonicity property that the standard screening models predict: if some type buys some pass, then every higher-type customer buys a (weakly) higher-priority pass. This monotonicity property does not generally hold in the setup of this paper because of the existence of externalities. There are two different intuitions on why monotonicity does not hold, which are illustrated by the following two examples.

Example 7 (Non-monotonicity 1)**.** Let $\big((N^h, N^l), K, (u^h, u^l)\big)$ be such that $N^h = N^l = 1$ and $K = 1$. Let $u^l = (4, 0)$ and $u^h = (5, 1)$ be the base utilities of the two types. Consider the scheme q with only the low-type customer in the queue. It can be verified that $p_1 = 4$ implements q and the high-type customer buys no pass. Calculations are shown in Figure 2.2. In this example, monotonicity fails because given the price, the second position is bad enough even for the high type and the high type would not want to join the queue.

Example 8 (Non-monotonicity 2)**.** Consider $\big((N^h, N^l), K, (u^h, u^l)\big)$ where $N^h = 2$, $N^l = 1$, and $K = 2$. Let $u^l = (6, 2, 1)$ and $u^h = (6.3, 2.2, 1.1)$ be the base utility functions of the two types. Consider the scheme $q = (q^h, q^l)$ such that $(q^h_0, q^h_1, q^h_2) = (0, 0, 2)$ and $(q^l_0, q^l_1, q^l_2) = (0, 1, 0)$. It can be verified that the price vector with $p_1 = 4.65$ and $p_2 = 1.65$ implements q. In this scheme, the high-type customers buy the lower-priority pass while the low-type customer buys the higher-priority pass. Calculations are shown in Figure 2.3. In this example, monotonicity fails because when a high type attempts to buy the higher-priority pass, the customer creates congestion to this pass, the externality of which can be larger than the utility of higher-priority customers when the base utility functions are convex and the two customer types are not very different.

By the definition of monotonicity, when the property does not hold, either a high type buys a strictly lower-priority pass, or the high type does not buy any pass at all. In Example 7, a low type buys some pass, but the high type does not buy any pass; in Example 8, the low type buys the higher-priority pass whereas the high type buys the lower-priority pass. Proposition 24 in the next subsection provides sufficient conditions under which some partial monotonicity holds.

Upper hemicontinuity of pure-strategy Nash equilibria with respect to the utility function means that when the two types are only slightly different, a two-pass scheme with two strictly concave base utilities is still not implementable. Intuitively, if the two types are sufficiently different, then we can separate the two types. The following proposition provides a cutoff of this closeness for a fixed scheme so that separation is impossible below the cutoff and possible above the cutoff.

Proposition 23 (Two-pass with two strictly concave utilities)**.** *Let* $((N^h, N^l), K, (u^h, u^l))$ *be such that* $K = 2$*. Assume* $u^h = u$ *and* $u^l = \beta^l u$*, where* u *is a strictly concave base utility function and* $\beta^l \in (0,1)$*. Let* $q \in \mathcal{Q}((N^h, N^l), 2)$ *be such that* $q_1^h = N^h$ *and* $q_2^h = N^l$*. The scheme* q *is implementable if and only if* $\beta^l \leq \frac{v(\theta_1)-v(\theta_2;\theta_1)}{v(\theta_1;\theta_2)-v(\theta_2)}$*, where* v *is the pass utility function constructed from* u*.*

We explain the result in relation to the standard screening models. Consider a scheme q such that $q_1^h = N^h$ and $q_2^l = N^l$. The implementability of q can be checked by the price vector p according to the formula[18]

$$p_2 = v^l(\theta_2), \quad p_1 = v^h(\theta_2) - \left[v^h(\theta_2;\theta_1) - p_2\right] = v^h(\theta_2) - \left[v^h(\theta_2;\theta_1) - v^l(\theta_2)\right],$$

where the term in the bracket on the right-hand side is the surplus given to the high-type customers. In this price vector, each high-type customer holding the high-priority pass is indifferent between the two passes, whereas each low-type customer holding the low-priority pass gets zero surplus. This property of the pricing formula is consistent with the optimal pricing formula in the standard screening models. However, the pricing formula in this paper has different compositions that lead to different results of implementability. For the pricing formula here,

$$p_1 - p_2 = v^h(\theta_1) - v^h(\theta_2;\theta_1) = v^h(\theta_1) - v^l(\theta_2) - \overbrace{\underbrace{\left(v^h(\theta_2;\theta_1) - v^h(\theta_2)\right)}_{\text{Downgrade externality}} - \underbrace{\left(v^h(\theta_2) - v^l(\theta_2)\right)}_{\substack{\text{Information rent in}\\ \text{standard screening}}}}^{\text{Total information rent}}, \tag{2.5}$$

where the second parenthesized difference is often called the information rent of the high-type customers in the standard screening models. In this paper, the additional term $v^h(\theta_2;\theta_1) - v^h(\theta_2)$ is the externality created when a high-type customer unilaterally downgrades to the low-priority pass. The externality is strictly positive and gives the higher-type customer higher total information rent, which further lowers p_1. To check whether q is implementable, IC_{21} of q implies

$$p_1 - p_2 \geq v^l(\theta_1;\theta_2) - v^l(\theta_2) = v^h(\theta_1) - v^l(\theta_2) - \left[\left(v^l(\theta_1) - v^l(\theta_1;\theta_2)\right) + \left(v^h(\theta_1) - v^l(\theta_1)\right)\right], \tag{2.6}$$

where the first parenthesized difference in the bracket is the externality created when the low-type customer upgrades to the high-priority pass. Since $v^h(\theta_1) - v^l(\theta_1) > v^h(\theta_2) - v^l(\theta_1) > 0$, in the standard screening models without externalities, the price difference from the pricing formula

[18] Lemma 10 in Appendix 2.A shows why the pricing formula can be used to check implementability.

would satisfy IC_{21}, making q implementable. However, from the proof of Proposition 18, when the base utility functions are strictly concave, $v^h(\theta_2;\theta_1)-v^h(\theta_2) > v^h(\theta_1)-v^h(\theta_1;\theta_2) > v^l(\theta_1)-v^l(\theta_1;\theta_2)$, which could still potentially make the price difference in (2.5) violate IC_{21}. When there is only one customer type, i.e.,

$$u^h = u^l \implies v^h(\theta_1) - v^l(\theta_1) = v^h(\theta_2) - v^l(\theta_2) = 0,$$

the right-hand side of (2.5) is strictly below the right-hand side of (2.6), making the scheme not implementable as in Proposition 18. When there are two different customer types, the proof of Proposition 23 shows that when the two customer types are sufficiently different, the price difference satisfies IC_{21}, making q implementable.

The condition for the implementation result above is about the two types being sufficiently different from each other. The next subsection shows that this intuition about different types being sufficiently different is still relevant as the model is extended to the general multi-type case.

2.5.2 General Multiple Utility Types

Suppose there are T types of basic utility functions. Like the two-type case, use superscripts to distinguish variables of different utility types and define other variables similarly. As in the two-type case, given a scheme q, for j, k, τ_1, and τ_2 such that $0 \leq j \leq K$, $1 \leq k \leq K$, and $1 \leq \tau_1 < \tau_2 \leq T$, $\text{IC}_{jk}^{\tau_2}$ implies $\text{IC}_{jk}^{\tau_1}$ if $0 < j < k$; $\text{IC}_{jk}^{\tau_1}$ implies $\text{IC}_{jk}^{\tau_2}$ if $j = 0$ or $1 \leq k < j \leq K$. Additionally, $\text{IR}_k^{\tau_2}$ implies $\text{IR}_k^{\tau_1}$. Let $\underline{t}_j = \max\{1 \leq \tau \leq T : q_j^\tau > 0\}$, which is the lowest type in θ_j, and $\bar{t}_j = \min\{1 \leq \tau \leq T : q_j^\tau > 0\}$, which is the highest type in θ_j. If $1 \leq j < k \leq K$, let $\text{IC}_{jk} = \text{IC}_{jk}^{\underline{t}_j}$; if $j = 0$, $1 \leq k \leq K$, and $\bar{t}_j \neq \emptyset$,[19] let $\text{IC}_{jk} = \text{IC}_{jk}^{\bar{t}_j}$; if $1 \leq k < j \leq K$, let $\text{IC}_{jk} = \text{IC}_{jk}^{\bar{t}_j}$. If $1 \leq k \leq K$, let $\text{IR}_k = \text{IR}_k^{\underline{t}_k}$.

Let the set of IC constraints be the collection of IC_{jk} over j and k such that $\sum_{t=1}^T q_j^t > 0$; let the set of IR collection of IR_k over k such that $1 \leq k \leq K$. Like the one-type case, implementation can be similarly defined with respect to the set of IC and IR constraints.

In Appendix 2.B, the lemmas for the two-type case are extended to the general multi-type case. An extension of Lemma 10 can be used to check the implementability of a scheme satisfying the restriction that $\underline{t}_j \leq \underline{t}_k$ if $1 \leq j < k \leq K$. Similar to the restriction introduced in the two-type case, the restriction for the multi-type case implies that every customer in a lower-priority pass cannot have a strictly higher type than does any customer in a higher-priority pass. It turns out that the restriction is part of a (partial) monotonicity property with concave base utility functions: if a customer of some type buys a pass, then every higher-type customer will buy a pass with a weakly higher priority in an implementable scheme.[20]

[19] The case where $\bar{t}_j = \emptyset$ is when no customer is in θ_j. By the definition of schemes, $\bar{t}_j = \emptyset$ only if $j = 0$.

[20] Recall that the monotonicity property does not hold in general, as Example 7 and Example 8 in the two-type case show.

Proposition 24 (Monotonicity with multiple concave utilities). *Fix* $((N^t)_{t=1}^T, K, (u^t)_{t=1}^T)$*, where* $K \geq 2$ *and* u^t *is concave for every* t*. Assume* $q \in ((N^t)_{t=1}^T, K)$ *is implementable. Fix* τ_1*,* τ_2*, and* l *such that* $1 \leq \tau_1 < \tau_2 \leq T$*,* $1 \leq l \leq K$*, and* $q_l^{\tau_2} > 0$*. If* $k > l$*, then* $q_k^{\tau_1} = 0$*. If* $\tau_1 \leq \underline{t}_{K-1}$*, then* $q_j^{\tau_1} > 0$ *for some* j *such that* $1 \leq j \leq l$*.*

The first part of the result says that in an implementable scheme, if a higher type buys some priority pass, then the pass cannot have a strictly lower priority than a lower-type customer does. The second part says that if a customer's type is also (weakly) higher than the lowest-type in the second-to-last priority pass, then this customer must purchase some (weakly) higher-priority pass than the lower-type customer does.[21]

When there are multiple customer types, implementing multi-pass schemes is possible if customers have utility functions that are sufficiently different from each other. The following result, a generalization of Theorem 5, characterizes the implementation conditions with respect to customer types.

Theorem 6 (Implementation with multiple linear utilities). *Fix* $((N^t)_{t=1}^T, K, (u^t)_{t=1}^T)$*. For every* t *and* n *such that* $1 \leq t \leq T$ *and* $1 \leq n \leq N$*, let* $u_n^t = \alpha^t - \beta^t nd$*, where* $d > 0$*. Assume* $\beta^1 > \beta^2 > \cdots > \beta^T = 1$ *and* $u_n^1 > u_n^2 > \cdots > u_n^T$ *for every* n *such that* $1 \leq n \leq N$*. Let* $q = \mathcal{Q}((N^t)_{t=1}^T, K)$ *be such that* $v^T(\theta_K) \geq 0$ *and* $q_0^t = 0$ *if* $1 \leq t \leq T$*. For each* k *such that* $2 \leq k \leq K$*, there exists* $b_k(\theta^{\underline{t}_1}, \ldots, \beta^{\underline{t}_{k-1}}; q) > 0$ *such that* q *is implementable if and only if the following two conditions hold:*

(a) $\underline{t}_j \leq \bar{t}_k$ *if* $1 \leq j < k \leq K$

(b) If $2 \leq k \leq K$*, then* $\beta^{\bar{t}_k} \leq b_k(\theta^{\underline{t}_1}, \ldots, \beta^{\underline{t}_{k-1}}; q)$*.*

Since a linear base utility function is concave, Part (a) of the result is immediate from Proposition 24. The inequality in Part (b) says that the slopes of different customer types need to be different enough. Similar intuitions can be applied to the strictly concave case. For the strictly concave case, focus on the case of $K = 2$, which is not implementable in the single-type case. When $K = 2$, implementation is possible for $\bar{t}_1 < \underline{t}_2$. Specifically, given a scheme q with $K = 2$ and strictly concave utility functions, by Proposition 24, it is necessary that $\bar{t}_1 < \underline{t}_2$. Given this necessary condition, the necessary and sufficient condition for the implementation of q is

$$v^{\underline{t}_1}(\theta_1) - v^{\underline{t}_1}(\theta_2; \theta_1) \geq v^{\bar{t}_2}(\theta_1; \theta_2) - v^{\bar{t}_2}(\theta_2),$$

which holds if $\underline{t}_1$ and $\bar{t}_2$ are sufficiently different. Thus for two-pass implementation, only the incentive constraints of two specific types matter. In other words, to implement a two-pass scheme where every customer type has at least one customer in the queue, it suffices to look for "gaps" between adjacent types.

For the family of strictly concave utility functions, the above notion of "gap" is not straightforward to illustrate. To simplify the intuition, consider a particular functional form where every pair of two functions from the family are an affine transformation of the other. The following result extends Proposition 23 and shows that two-pass implementation with multiple strictly concave utilities needs the lowest customer type in the first pass to be sufficiently different from the highest type in the second pass.

[21]Note that in Example 7, the condition is not met and the monotonicity property does not hold.

Proposition 25 (Two-pass implementation with multiple strictly concave utilities). *Let u be a strictly concave base utility function. Consider $((N^t)_{t=1}^T, K = 2, (u^t)_{t=1}^T)$, where $u^t = \alpha^t - \beta^t u$, $\beta^1 > \beta^2 > \cdots > \beta^T = 1$, and $u_n^1 > u_n^2 > \cdots > u_n^T$ for every n such that $1 \leq n \leq N$. Let $q \in \mathcal{Q}\left((N^t)_{t=1}^T, 2\right)$ be such that $v^T(\theta_2) \geq 0$ and $q_0^t = 0$ if $1 \leq t \leq T$. The scheme q is implementable if and only if*

$$\frac{\beta^{\underline{t}_1}}{\beta^{\bar{t}_2}} \geq \frac{v(\theta_1; \theta_2) - v(\theta_2)}{v(\theta_1) - v(\theta_2; \theta_1)}, \tag{2.7}$$

where v is the pass utility function constructed from u. In addition, the right-hand side of (2.7) is strictly greater than 1, and converges to 1 as $N^t \to \infty$ for every t such that $\bar{t}_2 \leq t \leq T$.

The inequality in the result is the requirement that the two adjacent customer types be sufficiently different. When the condition holds, the upper bound of $p_1 - p_2$ implied by the downward IC constraint is above the lower bound implied by the upward IC constraint, making the scheme implementable.

In summary, when there are multiple types of utility functions, the issue with resolving the upgrade and downgrade incentives is abated yet could persist. Specifically, when each customer type is an affine transformation of the other, implementing multi-pass schemes that are not implementable in the single-type case depends on how different (with respect to the slope) each pair of adjacent types are. In other words, the "gaps" between adjacent types need to large enough for implementation.

2.6 Discussions

In the single-type case, the one-pass scheme maximizes the profit, since by setting $p_1 = v(\theta_1)$, the park extracts all the surplus. In contrast, Section 2.6.1 revisits the two-type case and shows that a multi-pass scheme can be optimal. Section 2.6.2 discusses how the externality affects implementability. Section 2.6.3 discusses implementation in large queues where the size of externalities approaches zero. We also look at ϵ-implementation, where the customer does not switch to a different pass unless doing so improves the payoff by at least ϵ. Lastly, Section 2.6.4 looks at quantile queues, where the utility of positions in a queue depends on the relative quantiles in the queue.

2.6.1 Profits

In the standard screening models, provided that all customers are in the queue, having more passes weakly improves the profit because the park can extract even more surplus from those with a higher valuation. In contrast, for the single-type case in this paper, having more passes weakly hurts the profit: when $K = 1$ and $q_0 = 0$, setting $p_1 = v(\theta_1)$ implements $(q_0, q_1) = (0, N)$ and extracts all the customer surplus. On the other hand, implementing a multi-pass scheme means that the park needs to give away surplus to customers with higher-priority passes, making multi-pass scheme suboptimal with regard to profits. The critical intuition is that in the standard screening models, a single-pass scheme would not change the valuation of customers in the last pass. In contrast, in this paper, having all customers in one single pass increases the utility of those who used to be in the lowest-priority pass since now they have a probability of being at the front of the queue.

For the multi-type case, however, the single-pass scheme does not always dominate a multi-pass one. For the previously discussed two-type case with strictly concave utilities, the following proposition shows that for some values of β^l, the two-pass scheme is more profitable than a single-pass scheme. The proposition focuses on *all-serving* schemes, where every customer of every type buys some pass.

Proposition 26 (Profitability of two-pass schemes)**.** *Consider the two-type case and fix N^h and N^l. Let $u^h = u$, where u is strictly concave, and $u^l = \beta^l u$, where $\beta^l \in (0, 1)$. There exists $\bar{\beta} \in (0, 1)$ such that $K = 2$ is implementable, and the profit from the optimal all-serving two-pass scheme is higher than that of the optimal all-serving one-pass scheme if and only if $\beta^l \leq \bar{\beta}$.*

When the one-pass scheme serves both types, the park needs to respect the low type's IR constraint, lowering the price. If the two types are sufficiently different, then the park will have an incentive to add a higher-priority pass to extract more surplus from the high-type customer.

2.6.2 Implementation and Externality

In the standard screening models, every upward IC constraint holds if every local upward IC constraint holds, and every downward IC constraint hold if every local downward IC constraint holds. This constraint reduction does not hold in this paper. In the standard screening models, the crucial step to achieve this simplification for $1 < j < K$ would be $v(\theta_{j-1}; \theta_j) - v(\theta_j) \geq v(\theta_{j-1}; \theta_{j+1}) - v(\theta_j; \theta_{j+1})$, which does not hold in this paper as $v(\theta_{j-1}; \theta_j) = v(\theta_{j-1}; \theta_{j+1})$, i.e., the utility from switching only depends on whether the new pass has a higher or lower priority.

The lack of reduction of the upward IC constraints makes the implementation of multi-pass schemes non-trivial. In the standard screening models, the ID conditions imply that binding local downward IC constraints satisfies all upward IC constraints, and the local IC constraints together imply all the IC constraints. Hence the ID conditions are sufficient for implementation in the standard setup. However, in this paper, as the set of local upward IC constraints does not imply the set of upward IC constraints, the ID conditions do not guarantee implementation.

The IC constraints in this paper differ from the standard screening models in the existence of externality from switching: when one customer switches to another pass, the customer creates congestion if switching to a higher-priority pass and improves the waiting time if switching to a lower-priority pass. It turns out that the existence of this type of externality makes implementation harder than a model without externalities. We focus on the single-type case and formalize the externality with which relative implementability is discussed.

Definition 6. Fix (N, K) and $q \in \mathcal{Q}(N, K)$. A pass utility function v

(a) *creates externalities* if (2.1) holds.

(b) *creates more downgrade externalities* if v creates externalities and for every j and k such that $1 \leq j < k \leq K$, $0 < v(\theta_j) - v(\theta_j; \theta_k) < v(\theta_k; \theta_j) - v(\theta_k)$.

(c) *creates zero externality* if $v(\theta_j) = v(\theta_j; \theta_k)$ for every j and k such that $1 \le j < k \le K$.

By the definition of "constructed from u", for every base utility function u, the pass utility function v constructed from u creates externalities; furthermore, the proof of Proposition 18 shows that if u is strictly concave, v creates more downgrade externalities. When v creates zero externality, every scheme is implementable. In contrast, if v creates more downgrade externalities and $K \ge 2$, no scheme is implementable.

Proposition 27 (Externalities vs no externality)**.** *Fix* (N, K) *and* $q \in \mathcal{Q}(N, K)$.

(a) If v *creates zero externality, then* q *is implementable by setting* $p_k = v(\theta_k)$ *for each pass* θ_k.

(b) If v *creates more downgrading externalities and* $K \ge 2$*, then* q *is not implementable.*

By Part (a), since every pass is implementable when v creates zero externality, the set of implementable schemes when v creates externalities is a subset of the set when v creates zero externality. Moreover, Part (b) implies that the externalities can strictly shrink the set of implementable schemes. Specifically, when v creates more downgrade externalities, the externalities created by downgrading to a lower-priority pass are large relative to the externalities created by upgrading to a different pass, to the extent that the difference brings about an unresolvable conflict between incentivizing against upgrading and against downgrading when $K \ge 2$.

2.6.3 Implementation in Large Queues

This section analyzes how implementability changes when the total size of the queue grows. The interpretation here is that we are looking at a sequence of "markets" and see how implementability evolves with respect to N. The implementation results so far depend on the externalities customers impose on each other. By p^* in Lemma 7, a scheme fails to be implementable when the externalities from downgrading are too large, the price of a higher-priority pass has to be low to the extent that a customer in a lower-priority pass will have an incentive to upgrade. It turns out when the downgrade externalities can be sufficiently small with enough customers, which limits the price decrease of a higher-priority pass, then implementing schemes with many passes is possible.

Proposition 28 (Implementation with fixed K and large N)**.** *Fix* K *and a strictly decreasing sequence* $(u_n)_{n=1}^{\infty}$*. If* $\lim_{n\to\infty} \frac{u_n}{n} = 0$*, then there exist* M *and* $\underline{u} : \mathbb{N} \to \mathbb{R}$ *such that* $N \ge M$ *implies the existence of* $q \in \mathcal{Q}(N, K)$ *that is implementable with* $u_0 = \underline{u}(N)$*, i.e., there exists an implementable scheme for* (N, K, u) *when the reserve utility is* $\underline{u}(N)$.

The condition $\lim_n \frac{u_n}{n} = 0$ ensures that the local downgrade externality $v(\theta_k; \theta_{k-1}) - v(\theta_k)$ converges to 0 when $q_k \to \infty$ if $1 < k \le K$. The condition allows an unbounded utility function. For example, $u_n = -\log n$ is unbounded and $\lim_{n\to\infty} \frac{u_n}{n} = 0$. On the other hand, the condition implies a slow rate of decrease, excluding the linear and concave base utility functions.

If customers do not pay attention to small deviation gains, then the downward pressure on the price of a higher-priority pass is reduced. This inattention to small deviation gains can be formalized by the concept of ϵ-implementation, where each constraint of the park's optimization problem can be relaxed by an arbitrary and fixed $\epsilon > 0$. Specifically, IR_k in ϵ-implementation becomes

$$v(\theta_k) - p_k + \epsilon \geq u_0, \tag{ϵIR_k}$$

and IC_{jk} in ϵ-implementation becomes

$$v(\theta_j) - p_j + \epsilon \geq v(\theta_k; \theta_j) - p_k, \tag{ϵIC_{jk}}$$

and the definition of implementation is defined similarly to the exact-implementation case.

Definition 7 (ϵ-implementation)**.** Fix $N \geq K$.

(a) A scheme $q \in \mathcal{Q}(N, K)$ is ϵ-implementable if for each $\epsilon > 0$, there exists some price vector p such that ϵIR_k and ϵIC_{jk} hold for j and k such that $1 \leq j < k \leq K$.

(b) (N, K, u) is ϵ-implementable if there exists an ϵ-implementable scheme with N customers and K passes.

Let N grow and keep K fixed. Use the same condition $\lim_{n\to\infty} \frac{u_n}{n} = 0$ to ensure that local downgrade externalities are small with many customers. It turn out that in this case, with ϵ-implementation, the surplus-extracting price i.e., $p_k = v(\theta_k)$ for every k, ϵ-implements some schemes when N is allowed to grow.

Proposition 29 (ϵ-implementation)**.** *Fix K, $\epsilon > 0$, and a strictly decreasing sequence $(u_n)_{n=1}^{\infty}$. If $\lim_{n\to\infty} \frac{u_n}{n} = 0$, then there exist M and $\underline{u} : \mathbb{N} \to \mathbb{R}$ such that $N > M$ implies that some $q \in (N, K, u)$ is ϵ-implementable with $u_0 = \underline{u}(N)$ and leaves zero surplus to the customers, i.e., $p_j = v(\theta_j) - u_0$ for every j such that $1 \leq j \leq K$.*

To see how setting $p_k = v(\theta_k) - u_0$ implements a scheme, when $v(\theta_k; \theta_{k-1}) - v(\theta_k)$ is strictly smaller than ϵ, then $\epsilon\text{IC}_{k-1,k}$ holds with $p_k = v(\theta_k) - u_0$. The proof for the existence of M in the proposition statement is constructive: the cutoff M needs to be large enough so that the externality created from downgrading to a lower-priority pass is smaller than ϵ. This constructed cutoff is sufficient for ϵ-implementation yet not necessary. Particularly, given the assumptions, the higher-priority passes can have however many customers as long as there are much more customers in lower-priority passes.

2.6.4 Quantile Queues

So far, a customer's utility in a queue depends on the absolute position in the queue. Here we consider a special type of base utility function that depends on the relative position in a queue. Specifically, we look at quantile queues, where the utility of each position depends on the relative quantile of the position in a queue. Let u be a strictly decreasing and continuous linear function on $[0, 1]$, with $u(1) = 1$ and $u(0) = 0$. The utility of being in position n in a queue of N customers is $u(\frac{n-1}{N-1})$.

Assume there are T types of customers, each with m customers. Let $u^t = \beta^t u$ for some $\beta^t \geq 1$ be the base utility function of type t. Fix b and c such that $0 < b < c$, and assume β^t ranges uniformly from b to c as t ranges from 1 to T. Implementability of a particular type of scheme is analyzed: there are T priority passes, with every customer of the t-th type in the t-th priority pass.

Proposition 30 (Multi-Type quantile queues)**.** *Assume there are $T > 2$ types of customers, each of which has m customers and whose slopes are uniformly located on $[b, c]$ for some b and c such that $0 < b < c$. Let q be the scheme in which there are T passes and every customer of the t-th type is in the t-th priority pass. There exists $M(T) \in \mathbb{R}$ such that $M(T)$ is strictly increasing and linear in T and q is implementable if and only if $m \geq M(T)$.*

With fixed c and T, when there are sufficiently many customers in each priority pass, the externalities get small relative to the information rent as described in (2.5), making a scheme implementable. Specifically, the proof shows that $M(T) = \frac{b}{6(c-b)}(T-1) + \frac{1}{6}$, which is linear in T. With larger c, $M(T)$ grows more slowly in T. Figure 2.4 sets $b = 1$ and varies c and T to illustrate the evolution of $M(T)$ with respect to T and c. Fix the value of $M(T)$. Consider the T'-pass scheme in which each customer type has $M(T)$ customers and every customer of the t-th type is in the t-th priority pass. The scheme is implementable under (T', c) if (T', c) is on the left or on the curve of $M(T)$ and not implementable if strictly on the right. Since $\frac{c}{c-b} \geq 1$, for each $M(T)$, when T' is large enough, the scheme is not implementable for every c, and hence the vertical line at the tail of every curve.

2.7 Conclusions

This paper has shown the difficulty with implementing a multi-pass scheme under a static setting where customers make purchase decisions simultaneously and have uncertainty about the final position within each priority pass. The difficulty with implementing many passes derives directly from the conflict between incentivizing customers from upgrading and downgrading. When the base utility function is strictly concave, implementing a multi-pass scheme cannot be incentive compatible if there is only one type. This paper has shown show that such incentive conflicts can persist even when there are multiple types of base utility functions.

This paper uses a stylized model to deliver clear and concise insights that cover some important aspects of selling priority passes in a priority queue. Further enrichment to this paper's setup with respect to the paper's application merits further research. One suggested direction would be the combination of the simultaneous setup in this paper and the sequential setup in the queueing literature from operation research. Many parks sell priority passes as memberships with which a customer can enjoy the same priority for the customer's every visit in a given period. For example, if a customer buys a Platinum Season THE Flash Pass at Six Flags, the customer will enjoy the platinum priority for that customer's every visit to the park in that year. In such a scenario, if a customer has not purchased any priority pass upon the customer's first arrival at the park, the customer can observe the status of the queue at arrival time, which can affect the customer's pass purchase incentives. On the other hand, the purchase incentive is also going to be affected

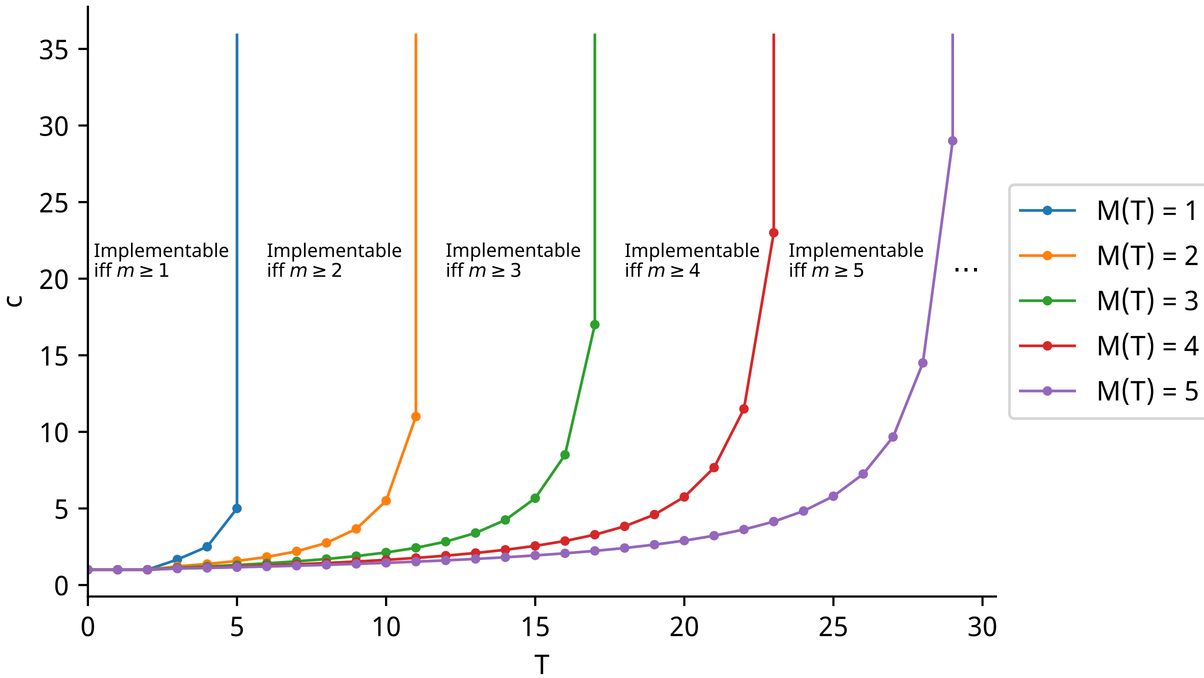

Figure 2.4: Evolution of $M(T)$ with respect to T and c, with $b = 1$. With fixed $M(T)$ and c, the T'-pass scheme in which every t-th type customer is in the t-th priority is implementable if the point (T', c) is on or to the left of the curve of $M(T)$; the scheme is not implementable if it is on the right.

by the forecast about the customer's future visits to the park. Both the number of future visits and the position in the queue for a future visit can be uncertain, which the simultaneous setup in this paper fits better. More complex modeling setups such as the one just described are out of the scope of this paper, and we leave the studies of these structures to future research.

2.A Lemmas for Two-Type Case

This section provides the intermediate results useful in proving the results in the two-type case. Figure 2.5 provides a roadmap of how results in this section contribute to the results in the two-type case.

The following two lemmas provide conditions under which the downward IC and IR constraint reductions hold in the two-type case.

Lemma 8 (IC Reduction with two utilities)**.** *Fix $((N^h, N^l), K, (u^h, u^l))$ with $K \geq 2$. Given a scheme q with a price vector p. Pick j and m with $1 \leq j < m \leq K$ and assume (p, q) satisfies $IC_{k,k+1}$ for every k such that $j \leq k < m$. If $q^l_j = 0$ or $q^l_k > 0$ for every k such that $j < k < m$, then (p, q) satisfies $IC_{j,m}$.*

Proof. Assume $q^l_j = 0$, which implies $\text{IC}_{j,m} = \text{IC}^h_{j,m}$. Note that by the construction of customer

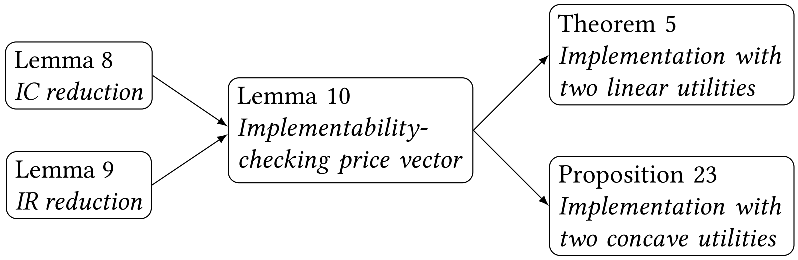

Figure 2.5: Roadmap of Appendix 2.A

types, for $j \le k < m$, $\mathrm{IC}_{k,k+1}$ implies $\mathrm{IC}^h_{k,k+1}$. Therefore, if $\mathrm{IC}_{k,k+1}$ holds for every k such that $j \le k < m$, then $\mathrm{IC}^m_{k,k+1}$ and thus $\mathrm{IC}^h_{j,m}$ holds by Lemma 6.

Now assume $q^l_k > 0$ for every k such that $j \le k < m$. By the construction of customer types, $\mathrm{IC}_{k,k+1} = \mathrm{IC}^l_{k,k+1}$ for every k such that $j \le k < m$, and thus IC_{jm} holds, again Lemma 6. □

Lemma 9 (IR Reduction with two utilities). *Fix $((N^h, N^l), K, (u^h, u^l))$ such that $K \ge 2$, $q \in \mathcal{Q}((N^h, N^l), K)$ and a price vector p. Assume for every j and k such that $1 \le j < k \le K$, (q, p) satisfies IC_{jk} and IR_k. If $q^l_j = 0$ or $q^l_k > 0$, then (q, p) satisfies IR_j.*

Proof. If $q^l_j = 0$, then $\mathrm{IR}_j = \mathrm{IR}^h_j$ and $\mathrm{IC}_{jk} = \mathrm{IC}^h_{jk}$. By IC_{jk} and the construction of customer types, $v^h(\theta_j) - p_j \ge v^h(\theta_k;\theta_j) - p_k \ge v^{\underline{t}_k}(\theta_k) - p_k \ge 0$, and thus IR_j holds.

If $q^l_k > 0$, then $\mathrm{IR}_k = \mathrm{IR}^l_k$. By IC_{jk} and the construction of types, $v^{\underline{t}_j}(\theta_j) - p_j \ge v^{\underline{t}_j}(\theta_k;\theta_j) - p_k \ge v^l(\theta_k;\theta_j) - p_k \ge 0$, and thus IR_j holds. □

The following result extends Lemma 7 in the single-type case and provides a pricing formula to check implementability.

Lemma 10 (Two-type implementation). *Fix $((N^h, N^l), K, (u^h, u^l))$ and $q \in \mathcal{Q}((N^h, N^l), K)$ where every customer buys some pass. Assume for j and k such that $1 \le j \le k \le K$, $q^l_j > 0$ implies $q^l_{j+1} > 0$. Let $p^* = (p^*_1, \ldots, p^*_K)$ such that $p^*_K = v^l(\theta_K)$ and $p^*_j - p^*_{j+1} = v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_{j+1};\theta_j)$. The scheme q is implementable if and only if (p^*, q) satisfies every upward IC constraint and $v^l(\theta_K) \ge 0$.*

Proof. Since $q^l_k > 0$ implies $q^l_{k+1} > 0$ for every k, $q^l_K > 0$ and hence $\mathrm{IR}_K = \mathrm{IR}^l_K$. Regardless of the implementability of q, (p^*, q) satisfies every local downward IC constraint. Pick j and k such that $1 \le j < k \le K$. If $q^l_j = 0$, then Lemma 8 implies IC_{jk} holds. If $q^l_j > 0$, then by assumption $q^l_r > 0$ for every r such that $j \le r < k$, and hence again by Lemma 8, IC_{jk} holds. Therefore, (p^*, q) satisfies every downward IC constraint. Since in addition IR^l_K holds, by Lemma 9, all the IR constraints also hold. Therefore, q is implementable if (p^*, q) satisfies every upward IC constraint.

Now assume q is implementable and let p implements q. Given $1 \le j < k \le K$, IC_{kj} implies $p_j - p_k \ge v^{\bar{t}_k}(\theta_j;\theta_k) - v^{\bar{t}_k}(\theta_k)$. On the other hand, $\mathrm{IC}_{l,l+1}$ for $j \le l < k$ implies

$$p_j - p_k = \sum_{l=j}^{k-1} p_l - p_{l+1} \le \sum_{l=j}^{k-1} v^{\underline{t}_l}(\theta_l) - v^{\underline{t}_l}(\theta_{l+1};\theta_l) = \sum_{l=j}^{k-1} p^*_l - p^*_{l+1} = p^*_j - p^*_k.$$

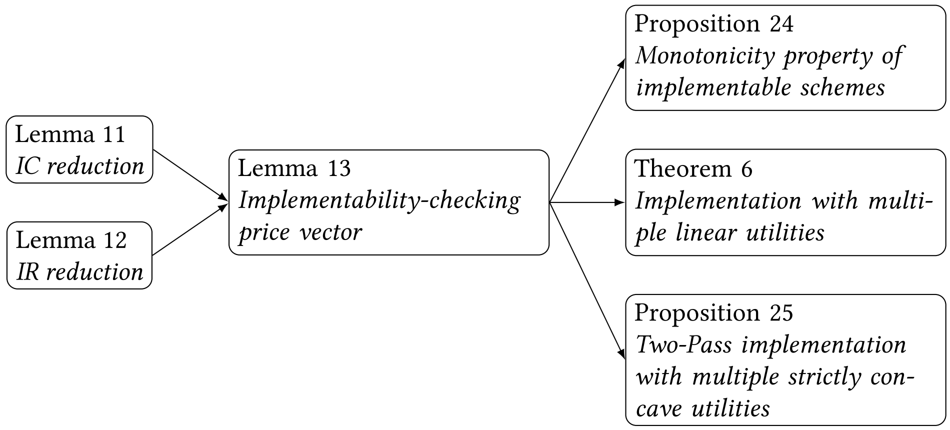

Figure 2.6: Roadmap of Appendix 2.B

Therefore, $p_j^* - p_k^* \geq p_j - p_k \geq v^{\bar{t}_k}(\theta_j; \theta_k) - v^{\bar{t}_k}(\theta_k)$, and hence (p^*, q) satisfies IC_{kj}. ☐

2.B Lemmas for Multi-Type Case

The lemmas in this section generalize the lemmas in the two-type case to the general multi-type case. A roadmap of how lemmas in this section contributes to the results in the multi-type case is illustrated in Figure 2.6.

Lemma 11 (Generalization of Lemma 8)**.** *Fix $\left((N^t)_{t=1}^T, K, (u^t)_{t=1}^T\right)$ with $K \geq 2$, a scheme $q \in \mathcal{Q}((N^t)_{t=1}^T, K)$, and a price vector p. Given j and l such that $1 \leq j < l$, assume $IC_{k,k+1}$ holds for every k such that $j \leq k < l$. If $\underline{t}_j \leq \underline{t}_k$ for every k such that $1 \leq j \leq k < l$, then IC_{jl} holds.*

Proof. Since $\underline{t}_j \leq \underline{t}_k$ for every k such that $j \leq k < l$, $\mathrm{IC}_{k,k+1}$ implies $\mathrm{IC}_{k,k+1}^{\underline{t}_j}$. Thus IC_{jl} holds by Lemma 6. ☐

Lemma 12 (Generalization of Lemma 9)**.** *Given $\left((N^t)_{t=1}^T, K, (u^t)_{t=1}^T\right)$ with $K \geq 2$, fix a scheme $q \in \mathcal{Q}((N^t)_{t=1}^T, K)$ and a price vector p. Assume there exists some $1 \leq j < k \leq K$ such that $\underline{t}_j \leq \underline{t}_k$. If IC_{jk} and IR_k hold, then IR_j holds.*

Proof. The proof is similar to that of Lemma 9 and hence the proof is omitted. ☐

The following result is an extension of Lemma 10 from the two-type case to the general multi-type case.

Lemma 13 (Implementation conditions in multi-type case)**.** *Fix $\left((N^t)_{t=1}^T, K, (u^t)_{t=1}^T\right)$. Let $q \in \mathcal{Q}((N^t)_{t=1}^T, K)$ be such that $\underline{t}_j \leq \underline{t}_k$ for j and k such that $1 \leq j < k \leq K$ and every customer buys some pass. Let $p^* = (p_1^*, \ldots, p_K^*)$ such that $p_K^* = v^T(\theta_K)$ and $p_j^* - p_{j+1}^* = v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_{j+1}; \theta_j)$. The scheme q is implementable if and only if (p^*, q) satisfies every upward IC constraint and $v^T(\theta_K) \geq 0$.*

Proof. Since every customer buys some pass and $\underline{t}_j \leq \underline{t}_k$ for j and k such that $1 \leq j \leq k \leq K$, $q_K^T > 0$ and IR_K^T holds. By Lemma 11 and Lemma 12, (p^*, q) satisfies every downward IC constraint and IR_j for j such that $1 \leq j \leq K$. Therefore, q is implementable if (p^*, q) satisfies every upward IC constraint. The rest of the proof is exactly the same with that of Lemma 10 □

2.C Proofs of Results in Main Text

2.C.1 Proof of Claim 1

Proof. When a customer downgrades from θ_{j_1} to θ_k, the range of positions is the same as when the customer downgrades from θ_{j_2} to θ_k. Therefore, $v(\theta_k; \theta_{j_1} = v(\theta_k; \theta_{j_2})$. With the same reasoning, $v(\theta_l; \theta_{l_1}) = v(\theta_k; \theta_{l_2})$.

Without any switching, positions of θ_k range from $Q_{k-1} + 1$ to Q_k. If a customer downgrades to θ_k from θ_{j_2}, then the range of positions of θ_k are now Q_{k-1} to Q_k. Since the front position of θ_k improves by one after the downgrade and the last position stays the same, $v(\theta_k; \theta_{j_2}) > v(\theta_k)$.

If a customer upgrades to θ_k from θ_{l_1}, then the positions of θ_k now range from $Q_{k-1} + 1$ to $Q_k + 1$. Since the front position of θ_k after the downgrade stays the same but the last position increases by one, $v(\theta_k) > v(\theta_k; \theta_{l_1})$. □

2.C.2 Proof of Claim 2

Proof. Let $a \in A$ be a strategy profile such that $\overline{q}(a) = q$ and fix a price vector p. We look at the incentives of customer i. If $a_i = \theta_0$, then the customer has no incentive to deviate to $\theta_j \neq \theta_0$ if and only if IC_{0j} holds.

Now assume $a_i = \theta_j \neq \theta_0$. The customer has no incentive to leave the queue if and only if IR_j holds. Lastly, for every k such that $1 \leq k \leq K$ and $k \neq j$, the customer has no incentive to deviate to θ_k if and only if IC_{jk} holds. Therefore, q is implementable if and only if there exists some price p such that (p, q) satisfies every constraint in the set of IR and IC constraints. □

2.C.3 Proof of Lemma 6

Proof. Fix $1 \leq j < k \leq K$. To prove Part (a), assume $\text{IC}_{l,l+1}$ for $j \leq l < K$. If $k = j + 1$, Part (a) is immediate. By induction, assume $\text{IC}_{j,k-1}$ and $\text{IC}_{k-1,k}$ hold, which implies

$$v(\theta_j) - p_j \geq v(\theta_{k-1}; \theta_j) - p_{k-1} \geq v(\theta_{k-1}) - p_{k-1} \geq v(\theta_k; \theta_{k-1}) - p_k = v(\theta_k; \theta_j) - p_k,$$

where the second and third inequalities come from $\text{IC}_{j,k-1}$ and $\text{IC}_{k,k-1}$; the first inequality and the last equality come from (2.1). Thus IC_{jk} holds.

To prove Part (b), assume IC_{jk} and IR_k hold, which implies

$$v(\theta_j) - p_j \geq v(\theta_j; \theta_k) - p_k \geq v(\theta_k) - p_k \geq 0,$$

where the inequalities come from (2.1), IC_{jk} and IR_k respectively. Thus IR_j holds.

Lastly, Part (c) is immediate when both IC_{jk} and IC_{kj} hold. □

2.C.4 Proof of Proposition 18

Proof. Assume $q \in \mathcal{Q}(N, K)$ is implementable. To arrive at a contradiction, it is without loss of generality to assume $N = Q_K$ because q is implementable with respect to (Q_K, N, u) if q is implementable with respect to (N, K, u). Since u is strictly concave and $N > 2$, there exists n such that $u_n - u_{n+1} < u_{n+1} - u_{n-2}$. If $q_k = 1$ for every k such that $1 \le j \le K$, then q is not implementable by Proposition 20 and the proof is complete. Therefore, assume we can pick j such that $1 \le j < K$ and $q_j > 1$ or $q_{j+1} > 1$.

For $m = 0, \dots, q_j$, define $x_m = u_{Q_j - m+1}$. Similarly, for $m = 0, \dots, q_{j+1}$, define $y_m = u_{Q_j + m}$. By construction, $y_{q_{k+1}} < \cdots < y_0 = x_0 < x_1 < \cdots < x_{q_j}$. Note $\mathrm{ID}_{j,j+1}$ implies

$$\begin{aligned} &\frac{\sum_{m=1}^{q_j} x_m}{q_j} - \frac{\sum_{m=0}^{q_{j+1}} y_m}{q_{j+1}+1} \ge \frac{\sum_{m=0}^{q_j} x_m}{q_j+1} - \frac{\sum_{m=1}^{q_{j+1}} y_m}{q_{j+1}} \\ \iff &\frac{\sum_{m=1}^{q_j}(x_m - x_0)}{q_j(q_j+1)} \ge \frac{\sum_{m=1}^{q_{j+1}}(y_0 - y_m)}{q_{j+1}(q_{j+1}+1)}. \end{aligned} \tag{2.8}$$

By the concavity of u, $\sum_{m=1}^{q_j}(x_m - x_0) \le \frac{q_j(q_j+1)(x_1-x_0)}{2}$, $\sum_{m=1}^{q_{j+1}}(y_0 - y_m) \ge \frac{q_{j+1}(q_{j+1}+1)(y_0-y_1)}{2}$, and $x_1 - x_0 \le y_0 - y_1$, which together imply

$$\frac{\sum_{m=1}^{q_j}(x_m - x_0)}{q_j(q_j+1)} \le \frac{x_1 - x_0}{2} \le \frac{y_0 - y_1}{2} \le \frac{\sum_{m=1}^{q_{j+1}}(y_0 - y_m)}{q_{j+1}(q_{j+1}+1)}.$$

The first inequality is strict if $q_j > 1$ and the lsat inequality is strict if $q_{j+1} > 1$. Thus at least one inequality must be strict, which violates $\mathrm{ID}_{j,j+1}$ and q is not implementable. □

2.C.5 Proof of Proposition 19

Proof. Let $q \in \mathcal{Q}(N, 2)$. For $n = 0, \dots, q_1$, let $x_n = u_{Q_1 - n+1}$. For $n = 0, \dots, q_2$, let $y_n = u_{Q_1 + n}$. By construction, $x_{q_1} > \cdots > x_0 = y_0 > \cdots y_{q_2}$. By convexity of u, $\sum_{n=1}^{q_1}(x_n - x_0) \ge \frac{q_1(q_1+1)(x_1-x_0)}{2}$, $\sum_{n=1}^{q_2}(y_0 - y_n) \le \frac{q_2(q_2+1)(y_0-y_1)}{2}$, and $x_1 - x_0 \ge y_0 - y_1$, which together imply

$$\frac{\sum_{m=1}^{q_1}(x_m - x_0)}{q_1(q_1+1)} \ge \frac{x_1 - x_0}{2} \ge \frac{y_0 - y_1}{2} \ge \frac{\sum_{m=1}^{q_2}(y_0 - y_m)}{q_2(q_2+1)},$$

where the two ends form a presentation of ID_{12} by (2.8).

It remains to show IC_{0k} holds for $k = 1, 2$. Consider $p_2 = v(\theta_2)$ and $p_1 = p_2 + v(\theta_1) - v(\theta_2; \theta_1)$. Since p binds IC_{12}, IC_{21} holds by ID_{12}. Since IR_2 is bound, IC_{02} holds. For IC_{01}, note that

$$v(\theta_1; \theta_2) - p_1 = v(\theta_1; \theta_2) - v(\theta_2) - v(\theta_1) + v(\theta_2; \theta_1) \le 0,$$

where the last inequality comes from ID_{12}. Thus IC_{01} holds and q is implementable. □

2.C.6 Proof of Lemma 7

Proof. By construction, (p^*, q) binds every local downward IC constraint and hence every downward IC constraint holds by Lemma 6. Since IR_K and every downward IC constraint hold, IR_k holds for k such that$1 \leq k \leq K$ by Lemma 6. Hence if (p^*, q) satisfies every upward IC constraint, p^* implements q.

Now assume q is implementable and let p be a price vector that implements q. For j and k such that $1 \leq j < k \leq K$, IC_{kj} implies $p_j - p_k \geq v(\theta_j; \theta_k) - v(\theta_k)$. On the other hand, $\text{IC}_{l,l+1}$ for $j \leq l < k$ implies

$$p_j - p_k = \sum_{l=j}^{k-1} p_l - p_{l+1} \leq \sum_{l=j}^{k-1} v(\theta_l) - v(\theta_{l+1}; \theta_l) = p_j^* - p_k^*.$$

Thus $p_j^* - p_k^* \geq p_j - p_k \geq v(\theta_j; \theta_k) - v(\theta_k)$, and (p^*, q) satisfies IC_{kj}. □

2.C.7 Proof of Proposition 20

Proof. Let $q \in \mathbb{Q}(N, K)$ where $q_k = 1$ for $1 \leq k \leq K$. Assume q is implementable. By p^* in Lemma 7,

$$p_1^* - p_3^* = \sum_{l=1}^{2} v(\theta_l) - v(\theta_{l+1}; \theta_l) = u_1 - \frac{u_1 + u_2}{2} + u_2 - \frac{u_2 + u_3}{2} = \frac{u_1 - u_3}{2}.$$

On the other hand, if $v(\theta_1; \theta_3) - v(\theta_3) = \frac{u_1 - u_2 - 2u_3}{2} > p_1^* - p_3^*$. Thus (p^*, q) does not satisfy IC_{31}, and by Lemma 7, q is not implementable. □

2.C.8 Proof of Proposition 21

Proof. If $K \leq 2$, by Proposition 19, every $q \in \mathbb{Q}(N, K)$ such that $v(\theta_K) \geq 0$ is implementable since u is convex.

Assume $K \geq 3$ and let $q \in \mathbb{Q}(N, K)$. It is without loss of generality to assume $N = Q_K$. Let $d = u_1 - u_2 > 0$. For j and k such that $1 \leq j < k \leq K$,

$$v(\theta_j) - v(\theta_j; \theta_k) = \frac{\sum_{n=Q_{j-1}+1}^{Q_j} u_n}{q_j} - \frac{\sum_{n=Q_{j-1}+1}^{Q_j+1} u_n}{q_j + 1} = \frac{\sum_{n=Q_{j-1}+1}^{Q_j} (u_n - u_{Q_j+1})}{q_j(q_j + 1)} = \frac{d}{2}.$$

Similarly, $v(\theta_k; \theta_j) - v(\theta_k) = \frac{d}{2}$. Therefore, by p^* in Lemma 7,

$$p_1^* - p_3^* = \sum_{l=1}^{2} v(\theta_l) - v(\theta_{l+1}; \theta_l) = v(\theta_1) - v(\theta_3) - \left[\sum_{l=1}^{2} v(\theta_{l+1}; \theta_l) - v(\theta_{l+1}) \right] = v(\theta_1) - v(\theta_3) - d.$$

On the other hand, $v(\theta_1) - v(\theta_1; \theta_3) = v(\theta_1) - v(\theta_3) - \frac{d}{2} > p_1^* - p_2^*$. Thus p^* does not satisfy IC_{31} and q is not implementable by Lemma 7. □

2.C.9 Proof of Theorem 5

Proof. By Proposition 21 and Proposition 22, when $K = 4$, it is necessary $q_1^l = q_2^l = 0$ and $q_3^h = q_4^h = 0$. By Lemma 10, it suffices to check whether p^* implements q. Since $v^l(\theta_K) \geq 0$, IR_K holds.

It remains to show that (p^*, q) satisfies every upward IC constraint. Begin with IC_{41}, which implies $p_1^* - p_4^* \geq v^l(\theta_1) - v^l(\theta_4) - \frac{d}{2}$. By the construction of p^*,

$$p_1^* - p_4^* = v^h(\theta_1) - v^h(\theta_2) + v^h(\theta_2) - v^h(\theta_3) + v^l(\theta_3) - v^l(\theta_4) - \beta^h d - \frac{d}{2},$$

and hence (p^*, q) satisfies IC_{41} if and only if $v^l(\theta_1; \theta_4) - v^l(\theta_4) \leq p_1^* - p_4^*$, i.e.,

$$v^h(\theta_1) - v^l(\theta_1) - \left[v^h(\theta_3) - v^l(\theta_3)\right] - \beta^h d \geq 0,$$

which implies $\beta^h \geq \frac{q_1+2q_2+q_3}{q_1+2q_2+q_3-2} > 1$. Similarly, IC_{42} implies $\beta^h \geq \frac{q_2+q_3}{q_2+q_3-1} > 1$ and IC_{31} implies $\beta^h \geq \frac{q_1+2q_2+q_3-1}{q_1+2q_2+q_3-2} > 1$. Hence there exists b such that q is implementable if and only if $\beta^h \geq b$. $\square$

2.C.10 Proof of Proposition 22

Proof. Fix $q \in \mathcal{Q}((N^h, N^l), K)$ such that for some $1 \leq j < k \leq K$, $q_j^l > 0$ and $q_k^h > 0$. By the construction of customer types, $\text{IC}_{jk} = \text{IC}_{jk}^l$ and $\text{IC}_{kj} = \text{IC}_{kj}^h$. Note that

$$v^l(\theta_j) - v^l(\theta_j; \theta_k) < v^h(\theta_j) - v^h(\theta_j; \theta_k) \leq v^h(\theta_k; \theta_j) - v^h(\theta_k),$$

where the first inequality comes from the construction of customer types and the second inequality comes from the concavity of v^h. However, the inequality from the two ends above violates ID_{jk}, and hence q is not implementable. $\square$

2.C.11 Proof of Proposition 23

Proof. With p^* in Lemma 10, IC_{21} implies

$$v(\theta_1) - v(\theta_2; \theta_1) = p_1^* - p_2^* \geq v^l(\theta_1; \theta_2) - v^l(\theta_2) = \beta^l \left[v(\theta_1; \theta_2) - v(\theta_2)\right],$$

which holds if and only if $\beta^l \leq \frac{v(\theta_1) - v(\theta_2;\theta_1)}{v(\theta_1;\theta_2) - v(\theta_2)}$. $\square$

2.C.12 Proof of Proposition 24

Proof. The two parts of the results are shown through the following two lemmas. The first lemma shows that if a higher-priority customer never buys a strictly lower-priority pass than a lower-type customer does in an implementable scheme; the second lemma shows that if the customer's type is (weakly) higher than the highest-type in θ_{K-1}, then this customer must necessarily buy some pass.

Lemma 14 (Higher-Priority for higher type). *Fix* $\left((N^t)_{t=1}^T, K, (u^t)_{t=1}^T\right)$ *and* $q \in \left((N^t)_{t=1}^T, K\right)$ *that is implementable. Assume* u^t *is concave for every* t*. If* $1 \le j < k \le K$*, then* $\underline{t}_j \le \bar{t}_k$*.*

Proof of lemma. Fix $q \in \mathcal{Q}((N^t)_{t=1}^T, K)$ where $\underline{t}_j > \bar{t}_k$ for some $1 \le j < k \le K$. It is without loss of generality to assume that every customer buys some pass. Note that

$$v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_j;\theta_k) < v^{\underline{t}_k}(\theta_j) - v^{\underline{t}_k}(\theta_j;\theta_k) \le v^{\underline{t}_k}(\theta_k;\theta_j) - v^{\underline{t}_k}(\theta_k),$$

where the first inequality comes from the assumption $\underline{t}_j > \bar{t}_K$ and the second inequality comes from the concavity of base utility functions. However, the inequality from the two ends violates ID_{jk} and hence q is not implementable. □

Lemma 15 (Pass-Buying with concave utilities). *Consider* $((N^t)_{t=1}^T, K, (u^t)_{t=1}^T)$ *where* $T \ge 2, K \ge 2$*, and* u^t *is concave for every* t*. Let* $q \in \mathcal{Q}((N^t)_{t=1}^T, K)$ *be implementable. Given* $1 \le j < K$ *and* $\tau \le \underline{t}_j$*, if* $\tau < \underline{t}_j$ *or* $\underline{t}_j < \underline{t}_K$*, then* $q_0^\tau = 0$*.*

Proof of lemma. By Lemma 14, $\underline{t}_k \le \underline{t}_K$. Fix p^* in Lemma 13. Note that

$$\begin{aligned} v^\tau(\theta_j;\theta_K) - p_j^* \ge v^{\underline{t}_j}(\theta_j;\theta_K) - p_j^* &= v^{\underline{t}_j}(\theta_j) - p_j^* - \left[v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_j;\theta_K)\right] \\ &\ge v^{\underline{t}_j}(\theta_K;\theta_j) - p_K^* - \left[v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_j;\theta_K)\right] \\ &= v^{\underline{t}_j}(\theta_K;\theta_j) - v^{\underline{t}_K}(\theta_K) - \left[v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_j;\theta_K)\right] \\ &\ge v^{\underline{t}_j}(\theta_K;\theta_j) - v^{\underline{t}_j}(\theta_K) - \left[v^{\underline{t}_j}(\theta_j) - v^{\underline{t}_j}(\theta_j;\theta_K)\right] \ge 0, \end{aligned}$$

where the first inequality comes from the assumption $\tau \le \underline{t}_j$, the second inequality comes from IC_{jK}, and the last equality comes from the concavity of base utility functions. If $\underline{t}_j = \underline{t}_K$, then $\tau < \underline{t}_j$ and the first inequality is strict; if $\underline{t}_j < \underline{t}_K$, then the third inequality is strict. Therefore, IC_{0j}^τ does not hold and hence $q_0^\tau = 0$. □

Let $q \in \mathcal{Q}((N^t)_{t=1}^T, K)$ be implementable. Fix τ_1, τ_2, and l such that $1 \le \tau_1 < \tau_2 \le T$, $1 \le l \le K$, and $q_l^{\tau_2} > 0$. If $q_k^{\tau_1} > 0$ for some $k > l$, then $\bar{t}_k < \underline{t}_l$ and q is not implementable by Lemma 14, a contradiction.

Suppose $\tau_1 \le \underline{t}_{K-1}$. Since u^t is concave for every t, $\underline{t}_{K-1} \le \bar{t}_K \le \underline{t}_K$ by Lemma 14. If $\underline{t}_{K-1} = \underline{t}_K$, then $\tau_1 < \tau_2 \le \underline{t}_K = \underline{t}_{K-1}$. In this case, by Lemma 15, $q_0^{\tau_1}$. If $\underline{t}_{K-1} < \underline{t}_K$, then by the same lemma, $q_0^{\tau_1} = 0$. Therefore, $q_0^{\tau_1} = 0$, and $q_j^{\tau_1} > 0$ for some j such that $1 \le j \le l$ by Lemma 14. □

2.C.13 Proof of Theorem 6

Proof. Since every customer buys some pass, Part (a) is immediate from Lemma 14. By p^* in Lemma 13, for j and l such that $1 \le j < l \le K$,

$$p_j^* - p_l^* = \sum_{k=j}^{l-1} p_k^* - p_{k+1}^* = \sum_{k=j}^{l-1} v^{\underline{t}_k}(\theta_k) - v^{\underline{t}_k}(\theta_{k+1}) - \beta^{\underline{t}_k}\frac{d}{2}.$$

Hence (p^*, q) satisfies IC_{lj} if and only if

$$p_j^* - p_l^* = \sum_{k=j}^{l-1} v^{\underline{t}_k}(\theta_k) - v^{\underline{t}_k}(\theta_{k+1}) - \beta^{\underline{t}_k}\frac{d}{2} \geq v^{\bar{t}_l}(\theta_j;\theta_l) - v^{\bar{t}_l}(\theta_k) = v^{\bar{t}_l}(\theta_j) - v^{\bar{t}_l}(\theta_l) - \beta^{\bar{t}_l}\frac{d}{2}.$$

Rearrange the terms to get

$$\left[v^{\underline{t}_j}(\theta_j) - v^{\bar{t}_l}(\theta_j)\right] - \left[v^{\underline{t}_{l-1}}(\theta_l) - v^{\bar{t}_l}(\theta_l)\right] - \left[\sum_{k=j}^{l-2} v^{\underline{t}_k}(\theta_{k+1}) - v^{\underline{t}_{k+1}}(\theta_{k+1})\right] + \frac{d}{2}\left[\beta^{\bar{t}_l} - \sum_{k=j}^{l-1}\beta^{\underline{t}_k}\right] \geq 0.$$

Plug the utility functions $u_n^t = \alpha^t - \beta^t nd$ in the inequality above to get

$$\begin{aligned}&\left[\alpha^{\underline{t}_j} - \alpha^{\bar{t}_l} - (\beta^{\underline{t}_j} - \beta^{\bar{t}_l})Q_{j-1}d - \frac{(\beta^{\underline{t}_j} - \beta^{\bar{t}_l})(1+q_j)}{2}d\right] \\ &\qquad - \left[\alpha^{\underline{t}_{l-1}} - \alpha^{\bar{t}_l} - (\beta^{\underline{t}_{l-1}} - \beta^{\underline{t}_l})Q_{l-1}d - \frac{(\beta^{\underline{t}_{l-1}} - \beta^{\bar{t}_l})(1+q_l)}{2}d\right] \\ &- \left[\sum_{k=j}^{l-2}\alpha^{\underline{t}_k} - \alpha^{\underline{t}_{k+1}} - (\beta^{\underline{t}_k} - \beta^{\underline{t}_{k+1}})Q_k d - \frac{(\beta^{\underline{t}_k} - \beta^{\underline{t}_{k+1}})(1+q_{k+1})}{2}d\right] + \frac{d}{2}\left[\beta^{\bar{t}_l} - \sum_{k=j}^{l-1}\beta^{\underline{t}_k}\right] \geq 0,\end{aligned}$$

from which $\beta^{\bar{t}_l}$ can be solved for to get

$$\beta^{\bar{t}_l} \leq \sum_{k=j}^{l-1}\frac{-1 + q_k + q_{k+1}}{-1 + q_j + q_l + 2\sum_{m=j+1}^{l-1} q_m}\beta^{\underline{t}_k}.$$

Let $\beta_j^{\bar{t}_l}$ be the right-hand side of the inequality above, which is independent of d. For $2 \leq l \leq K$, let $b_l(\beta^{\underline{t}_1}, \dots, \beta^{\underline{t}_{k-1}}) = \min_{j<l}\beta_j^{\bar{t}_l}$. Conditional on Part (a), q is implementable if and only if $\beta^{\bar{t}_l} \leq b_l(\beta^{\underline{t}_1}, \dots, \beta^{\underline{t}_{k-1}})$ for every l such that $2 \leq l \leq K$. □

2.C.14 Proof of Proposition 25

Proof. Since every customer buys some pass and each u^t is strictly concave, by Proposition 18 and Lemma 14, $\underline{t}_1 = \bar{t}_2 - 1$. Since $K = 2$, ID_{12} is sufficient and necessary for implementation, which implies

$$\frac{\beta^{\underline{t}_1}}{\beta^{\bar{t}_2}} \geq \frac{v(\theta_1;\theta_2) - v(\theta_2)}{v(\theta_1) - v(\theta_2;\theta_1)} > 1,$$

where the second inequality comes from the strict concavity of u. If $N^t \to \infty$ for every τ such that $\bar{t}_2 \leq \tau \leq T$, then q_2 tends to infinity, and hence both $v(\theta_2)$ and $(\theta_2;\theta_1)$ converge to $-\infty$. Thus the right-hand side converges to 1. □

2.C.15 Proof of Proposition 26

Proof. By Proposition 23, a two-pass all-serving implementable scheme exists if and only if $\beta^l \leq \frac{v(\theta_1)-v(\theta_2;\theta_1)}{v(\theta_1;\theta_2)-v(\theta_2)}$. By p^* in Lemma 10, the revenue of the all-serving two-pass scheme is

$$N^l p_2^* + N^h p_1^* = N^l \beta^l v(\theta_2) + N^h \left[\beta^l v(\theta_2) + \beta^l (v(\theta_1;\theta_2) - v(\theta_2))\right].$$

The revenue of the all-serving one-pass scheme is $\beta^l(N^h v(\theta_1) + N^l v(\theta_2))$. Thus conditional on $v(\theta_1) - v(\theta_2;\theta_1) \geq \beta\left[v(\theta_1;\theta_2) - v(\theta_2)\right]$, the two-pass scheme is better than the one-pass scheme if and only if

$$N^l \beta^l v(\theta_2) + N^h \left[\beta^l v(\theta_2) + v(\theta_1) - v(\theta_2;\theta_1)\right] \geq \beta^l \left(N^l v\left(\theta_2\right) + N^h v(\theta_1)\right).$$

The inequality implies

$$\beta^l \leq \frac{v(\theta_1) - v(\theta_2;\theta_1)}{v(\theta_1) - v(\theta_2)} \leq \frac{v(\theta_1) - v(\theta_2;\theta_1)}{v(\theta_1;\theta_2) - v(\theta_2)},$$

where the right-hand side is in $(0, 1)$. To conclude, let $\overline{\beta} = \frac{v(\theta_1)-v(\theta_2;\theta_1)}{v(\theta_1)-v(\theta_2)}$. □

2.C.16 Proof of Proposition 27

Proof. Assume v creates zero externality. Consider the price vector $p_k = v(\theta_k)$ for every pass k. For every pass j, IR_j holds. Additionally, since $v(\theta_k;\theta_j) - p_k = v(\theta_k;\theta_j) - v(\theta_k) = 0$, IC_{jk} holds. Thus p implements q when v creates zero externality.

Now assume $K \geq 2$ and v creates more externalities. Given $q \in \mathcal{Q}(N, K)$, to arrive at a contradiction, assume p implements q. Thus IC_{12} and IC_{21} imply

$$v(\theta_1;\theta_2) - v(\theta_2) \leq p_1 - p_2 \leq v(\theta_1) - v(\theta_2;\theta_1).$$

However, as v creates more downgrade externalities, $v(\theta_1) - v(\theta_1;\theta_2) < v(\theta_2;\theta_1) - v(\theta_2)$, and hence IC_{12} and IC_{21} cannot both hold, a contradiction. □

2.C.17 Proof of Proposition 28

Proof. The proof focuses on schemes with every customers buying some pass. To construct a K-pass implementable scheme, first fix $q_1 > 0$. By p^* in Lemma 7, IC_{21} holds if

$$v(\theta_2;\theta_1) - v(\theta_2) \leq v(\theta_1) - v(\theta_1;\theta_2).$$

The right-hand side is strictly positive, and thus for IC_{21} to hold, it is sufficient for the left-hand side to converge to 0 as q_2 grows. To see this, note that for every k such that $1 < k \leq K$,

$$v(\theta_k;\theta_{k-1}) - v(\theta_k) = \frac{u_{Q_{k-1}}}{q_k + 1} - \frac{v(\theta_k)}{q_k + 1}.$$

The first term, $\frac{u_{Q_{k-1}}}{q_k+1}$, converges to 0 as $q_k \to \infty$. For the second term, note that

$$\lim_{q_k\to\infty} \frac{v(\theta_k)}{q_k+1} = \lim_{q_k\to\infty} \frac{1}{q_k+1} \frac{\sum_{n=Q_{k-1}+1}^{Q_k} u_n}{q_k} = 0,$$

where the last equality holds because $\lim_{n\to\infty} \frac{u_n}{n} = 0$. Therefore, for fixed q_1, IC_{21} holds strictly with large q_2. Suppose that with this procedure, we have picked $q_1, q_2, \dots, q_{l-1}$ for some $2 < l \leq K$ such that IC_{kj} holds strictly for every j and k such that $1 \leq j < k < l$. Since $\text{IC}_{l-1,j}$ holds strictly for every j such that $1 \leq j < l-1$,

$$\begin{aligned} v(\theta_j) - v(\theta_{l-1}) - \left[v(\theta_j) - v(\theta_j;\theta_{l-1})\right] &= v(\theta_j;\theta_{l-1}) - v(\theta_{l-1}) \\ &< \sum_{k=j}^{l-2} v(\theta_k) - v(\theta_{k+1};\theta_k) \\ &= v(\theta_j) - v(\theta_{l-1}) - \sum_{k=j+1}^{l-1} \left[v(\theta_k;\theta_{k+1}) - v(\theta_k)\right], \end{aligned}$$

and thus

$$v(\theta_j) - v(\theta_j;\theta_{l-1}) > \sum_{k=j+1}^{l-1} \left[v(\theta_k;\theta_{k+1})\right].$$

Now similarly for IC_{lj} to hold with p^*, it is necessary that

$$v(\theta_j) - v(\theta_j;\theta_{l-1}) \leq v(\theta_l;\theta_{l-1}) - v(\theta_l) + \sum_{k=j+1}^{l-1} \left[v(\theta_k;\theta_{k+1}) - v(\theta_k)\right],$$

which holds strictly with large q_l since we have shown that $v(\theta_k;\theta_{k-1}) - v(\theta_k)$ converges to 0 as $q_k \to \infty$ for every k such that $1 < k \leq K$. Therefore, there exists large enough q_l such that IC_{kj} holds with p^* for every j and k such that $1 \leq j < k \leq l$. Repeat this procedure for all $1 < l \leq K$, and we have constructed an implementable scheme. Therefore, there exists $M > 0$ such that if $N > M$ then there exists $q \in \mathcal{Q}(N, K)$ that is implementable. □

2.C.18 Proof of Proposition 29

Proof. If there exists some $q \in \mathcal{Q}(N, K)$ and p such that (p, q) that satisfies every IC constraint, then setting $\underline{u}(N) = v(\theta_K) - p$ makes q implements q with $u_0 = \underline{u}(N)$. Fix $q \in \mathcal{Q}(N, K)$ where every customer buys some pass. For every k such that $1 \leq k \leq K$, let $p_k = v(\theta_k)$. With this construction, ϵIC_{kj} holds for every j and k such that $1 \leq j < k \leq K$. For ϵIC_{jk} to hold, it is necessary that

$$v(\theta_k;\theta_j) - p_k - \left[v(\theta_j) - p_j\right] - \epsilon = \frac{u_{Q_{k-1}}}{q_k+1} - \frac{v(\theta_k)}{q_k+1} - \epsilon < 0.$$

Note that $\lim_{q_k\to\infty}\frac{u_{Q_{k-1}}}{q_k+1}=0$. It remains to show that $\lim_{q_k\to\infty}\frac{v(\theta_k)}{q_k+1}=0$, which can be shown to be implied by $\lim_{n\to\infty}\frac{u_n}{n}=0$. To see this, note that $\frac{v(\theta_k)}{q_k+1}=\frac{1}{q_k+1}\sum_{n=Q_{k-1}+1}^{Q_k}\frac{u_n}{q_k}$, which would converge to 0 when q_k tends to infinity if $\lim_{n\to\infty}\frac{u_n}{n}=0$. Hence ϵIC_{jk} holds. Therefore, given $q_1>0$, there exists q_2 large enough such that IC_{12} holds with $p_k=v(\theta_k)$ for every k such that $1\le k\le K$. Then given $q_1,\ldots,q_{k-1}$ for some $k\le K$, there exists q_k large enough such that IC_{jk} holds for every j such that $1\le j\le k-1$. The procedure terminates after a finite number of times since K is fixed. Thus there exists M such that for all $N\ge M$, there is a scheme $(q_1,\ldots,q_K)$ that can be implemented by $p_k=v(\theta_k)$ for every K such that $1\le k\le K$. □

2.C.19 Proof of Proposition 30

Proof. Let $b=\beta^1>\cdots>\beta^T=c$ be the T customer types. Fix j and k such that $1\le j<k\le K=T$. Given the construction of base utility functions in this setup, $v^t(\theta_j)-v^t(\theta_k)=\frac{\beta^k(k-j)m}{mT-1}$. With p^* in Lemma 13, IC_{kj} implies $\sum_{t=j}^{k-1}v^t(\theta_t)-v^t(\theta_{t+1};\theta_t)\ge v^k(\theta_j;\theta_k)-v^k(\theta_k)$, which is equivalent to

$$\sum_{t=j}^{k-1}\frac{\beta^t m}{mT-1}-\frac{\beta^t}{2(mT-1)}\ge\frac{\beta^k(k-j)m}{mT-1}-\frac{\beta^k}{2(mT-1)}.$$

Multiply both sides by $2(mT-1)$ and rearrange the terms to get

$$2m\ge 1+\frac{\beta^k(k-j-1)}{\sum_{t=j}^{k-1}(\beta^t-\beta^k)}.$$

Thus IC_{kj} holds if the inequality above holds. For fixed j and k such that $1<j<k\le K$, the right-hand side of the inequality is larger if both j and k decrease by 1, because the denominator would be the same but $\beta^{k-1}>\beta^k$. Therefore, to find the pair of j and k such that the right-hand side is maximized, it is sufficient to pick from j and k such that $j=1$. For each $2<k\le T$, set

$$m_k(T)=\frac{1}{2}+\frac{\beta^k(k-2)}{2\sum_{l=1}^{k-1}(\beta^l-\beta^k)}=\frac{1}{2}+\frac{\beta^k(T-1)}{c-b}\times\frac{k-2}{k(k-1)}. \tag{2.9}$$

Note that $\max_{k\ge2}m_k(T)$ can be explicitly expressed. To see this, since $\frac{d}{dk}\left(\frac{k-2}{k(k-1)}\right)=\frac{-(k^2-4k+2)}{k^2(k-1)^2}$, the derivative is positive for $k\in(2-\sqrt{2},2+\sqrt{2})$ and negative otherwise. Therefore, for integers $k>2$, the right-hand side of (2.9) is maximized at either $k=3$ or $k=4$. It turns out that $\frac{k-2}{k(k-1)}=\frac{1}{6}$ for both $k=3$ and $k=4$. Since β^k is decreasing in k , (2.9) is maximized at $k=3$. Let $M(T)=m_3(T)$, i.e.,

$$M(T)=\frac{1}{2}+\frac{\beta^3(T-1)}{6(c-b)}=\frac{c}{6(c-b)}(T-1)+\frac{1}{6}.$$

Since (p^*,q) satisfies every IC constraint if and only if $m\ge M(T)$, q is implementable if and only if $m\ge M(T)$. Lastly, observe that $M(T)$ is strictly increasing in T and $M(\infty)=\infty$. □

Chapter 3

A Model of Sponsored Reviews

3.1 Introduction

Thanks to the popularity of internet platforms such as YouTube and TikTok, more and more consumers are following internet bloggers for the bloggers' reviews of new products. That the bloggers have many subscribers that rely on their reviews for purchase decision makes bloggers a potential sales channel. This type of marketing is often called influencer marketing. To incentivize the bloggers to help promote a firm's product, one common incentive is for the firm to sponsor a blogger's product review by offering sales commissions for purchases by the blogger's subscribers. In a sponsored review, since the blogger now benefits from higher sales, the blogger now has a bigger incentive to review the product favourably, even though the blogger's own private signal about the product's quality says otherwise. On the other hand, the blogger cares about the accuracy of the review. The motivation for the desire for accuracy could be both internal and external: an internal motivation could be the blogger's self-esteem from writing accurate reviews; an external motivation could be the understanding that accurate reviews attract more viewership, for which the internet platform on which the blogger publishes reviews rewards the blogger. The research question in this question is under what conditions it would be possible for the blogger to provide an accurate review.

The main finding of the paper is that whether a blogger has an incentive to provide a honest review depends crucially on how informativeness the blogger's private signal about the reviewed product's quality is. The more informative the blogger's signal is, the more likely that the in a biased review, the review will turn out to be inaccurate, which disciplines the blogger to provide an honest review. Given the intuition about signal informativeness, the paper finds that with regard to the prior belief about a reviewed product's quality, the blogger has an incentive to truthfully review if the prior is within some middle range of quality priors because it turns out the blogger's signal is most informative within this middle range. The paper is to show that this middle range of priors in which the incentive for an honest review is the strongest is widened if the blogger has higher signal precision or the sales commission rate offered by the firm is lower. Lastly, the paper finds that the existence of the sponsorship, in the form of sales commissions, also creates a commitment problem to the blogger in face of multiple equilibria: if the blogger can commit to an equilibrium before the signal arrives, the blogger will still prefer to commit to truthfully reviewing the product even when the review is sponsored.

For the blogger's subscribers, when the blogger offers a sponsored review, it is in their interest to work out whether the blogger is being honest about the review, since now the blogger can also profit from leading the subscribers to belief the reviewed product is good with a high probability. For example, when the blogger says that the reviewed product is above expectation, the blogger needs to figure out how likely that the product is actually below expectation but the blogger inflates the probability of the quality being good. When the subscribers have no way to verify the blogger's message, talk is cheap. Indeed, this paper answers the research question by modelling a specialized cheap-talk game, which Crawford and Sobel (1982) provides a foundational framework.[1] Specifically, in this paper, the blogger communicates with the subscribers through costless and unverifiable messages to the subscribers to try to lead the subscribers to form an intended belief. In contrast to that paper, in this paper, the blogger, which is the sender in that paper, has incomplete yet useful information about the state. Moreover, in this paper, the blogger's signal structure is finitely discrete, which enables us to provide sharper prediction about when the blogger has an incentive to always truthfully reveals the blogger's information to the subscribers.

With regard to sponsored blogs and reviews, Hwang and Jeong (2016) conducts an empirical analysis on consumer perception of different forms of sponsor disclosure. The consideration about sponsorship disclosure is moot in this paper because this paper makes the assumption that the subscribers have no way to verify whether the blogger has taken a sponsorship for the review. Since having a sponsorship brings extra income to the blogger and a blogger has no way to show to the subscribers that the review is not sponsored, this paper is to show that in equilibrium, the blogger always takes up the sponsorship from the firm.

There are a number of theoretical studies in product reviews. For example, Yubo Chen and Xie (2005) looks into how a firm can adapt its marketing strategy to product reviews made by third-parties such as magazines. A later study by the same authors analyse how the firm can adjust its marketing mix to address several strategic issues related to online consumer reviews. In contrast to these papers, this paper focuses on the strategic consideration of the reviewers themselves. In the two papers cited here, the blogger's strategic consideration is taken as given and these papers study how the firm reacts to these reviews. In contrast, this paper takes as given the reaction of the firm to the blogger's review, such as the design of the sponsorship contract. Instead, this paper studies the blogger's reviewing incentive with the offer of a sponsorship.

Section 3.2 constructs the model of sponsored review, with results and some discussions about the results presented in Section 3.3. Lastly, Section 3.4 concludes the paper with some comments on future research.

3.2 Model

A blogger has a finitely sized continuum of subscribers each of whom has type $\mu \in [0, 1]$, which is the lowest probability of a product being good that the subscriber of this type is willing to purchase the product. Assume the distribution of the subscribers' types follows a continuous

[1]See Farrell and Rabin (1996) for a nice survey and discussion of cheap talk.

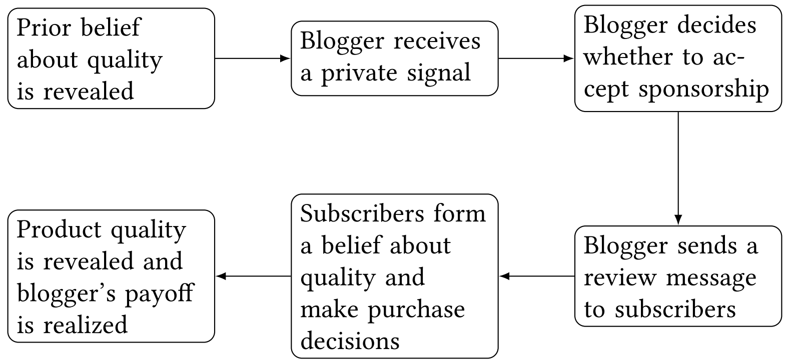

Figure 3.1: Timing of the model

distribution whose cumulative distribution function is $F : [0,1] \to [0,1]$. The blogger samples a product to review. Let $\theta \in \{0,1\}$ be the quality of a product, with 1 meaning that the product quality is good and 0 indicating a bad quality. Let $\lambda \in [0,1]$ be the prior probability of the product being good and assume the prior belief is common knowledge. After sampling the product, the blogger receives a costless private signal $s \in \{0,1\}$ such that $\Pr(s = i|\theta = i) = r \in (1/2, 1)$, i.e., the precision of the blogger's signal is r.[2] Assume the blogger's signal structure is common knowledge, i.e., each subscriber knows the distribution of the blogger's signal conditional on the product's quality.

Upon receiving the signal of a product with prior λ, the blogger updates the belief about θ using Bayes rule, and then makes a decision on whether to take up a sponsorship for the review. If he does not take a sponsorship, he only makes money from the platform on which he publishes his reviews. If the reviewer accepts the sponsorship, in addition to the income from the platform, the reviewer receives sales commission from the firm for each purchase by the blogger's subscribers. We assume that subscribers cannot observe nor verify whether the blogger has taken a sponsorship.

After the decision on whether to take the sponsorship, if the blogger's realized signal is s, the blogger sends his subscribers a message $m \in \mathscr{M}(s) \subseteq \mathbb{R}$, where $\mathscr{M}(s)$ is the message space when the realized signal is s.[3] Assume $\mathscr{M}(s) \neq \emptyset$ for $s \in \{0,1\}$. Having received the blogger's message, the subscribers form a belief about the product's quality and then make purchase decision based on the belief and each subscriber's type. Assume $\mathscr{M}(1) \subseteq \mathscr{M}(0)$, i.e., when the signal is low, the blogger can replicate any message sent when the signal is high. With some abuse of notation, denote the aggregate message space by $\mathscr{M} = \mathscr{M}(0) \cup \mathscr{M}(1)$ and assume $\mathscr{M}$ has at least two elements.[4] Figure 3.1 draws the timing of the model.

[2] The assumption that $r > 1/2$ is without loss of generality since if $r < 1/2$, then the blogger could simply use the alternative signal $\tilde{s}$ such that $\tilde{s} = 0$ when $s = 1$ and $\tilde{s} = 1$ when $s = 0$.

[3] That the messaging space does not depend on whether the blogger takes up a sponsor is consistent with the assumption that sponsorship is neither observable nor verifiable.

[4] For example, the message space could consist of only two messages that claim to have received the high or low signal.

3.2.1 Subscribers' strategy

Each subscriber makes a decision on whether to purchase the reviewed product after receiving hte blogger's message. Given the prior λ, let $b(\cdot;\lambda) : \mathscr{M} \to [0,1]$ be a **belief mapping** that maps a message to a probability of the quality being good. If a subscriber with type $\mu \in (0,1)$ has belief $b \in (0,1)$ after receiving the blogger's message, then the subscriber purchases the reviewed product if and only if $\mu \leq b$. Since the blogger's signal structure is common knowledge, the subscribers know that for a product with prior λ, the blogger's posterior probability of the product being good is $q_1(\lambda) = \frac{\lambda r}{\lambda r + (1-\lambda)(1-r)}$ when $s = 1$ and $q_0(\lambda) = \frac{\lambda(1-r)}{\lambda(1-r)+(1-\lambda)r}$ when $s = 0$. Therefore, for a subscriber to form a belief about the product's quality after receiving the message, the subscriber first forms a belief about which signal the blogger got, and then updates the belief about quality using Bayes rule. In equilibrium, which is to be formally defined later, we require the each subscriber's belief is consistent with the blogger's strategy using Bayes rule.

3.2.2 Blogger's strategy

Given a prior μ and realized signal s, the blogger's strategy is $\Phi(s;\lambda) \in \Delta(\mathscr{M}(s))$, where $\Delta(\mathscr{M}(s))$ is the set of probability measures on $\mathscr{M}(s)$. With some abuse of notation, let $\Phi(\lambda) = \Phi(0;\lambda)\Pr(s = 0|\lambda) + \Phi(1;\lambda)\Pr(s = 1|\lambda)$ be the unconditional message distribution given the blogger's strategy. One regularity condition we impose on the blogger's strategy throughout the paper is that for every $m \in \text{supp}(\Phi(\lambda))$, $\Pr(s = 1|m, \{\Phi(\tilde{s};\lambda)\}_{\tilde{s}})$ is defined.[5] After the subscribers have made their decisions, the state, which is the quality of the product, is revealed to everyone. The platform on which the blogger posts reviews rewards the blogger for posting accurate reviews, i.e., where an accurate review is where the subscribers are led to a belief that is close to the revealed state. An interpretation of this assumption under the paper's context is that good reviews tend to have more viewership, which translates into more traffic to the platform, and hence the website is willing to reward the blogger for an accurate review. In addition, if the blogger takes the firm's sponsorship, the firm rewards the blogger for bringing revenue to the firm. If b is the subscribers' belief after the blogger's message, then the blogger's utility, conditional on the acceptance of the sponsorship, is

$$u(b,\theta) = M\left[C - (b-\theta)^2 + \alpha F(b)\right], \tag{3.1}$$

where M denotes the size of the subscribers. Assume $M = 1$. The second term $(b-\theta)^2$ measures the disutility of deviating from the revealed state, which can be interpreted both through the platform's rewards and the blogger's interval desire for accurate reviews. The interpretation of the assumption that the blogger's utility is decreasing in the deviation from the revealed state is that inaccurate reviews attract smaller traffic than an accurate one, and hence the platform rewards the blogger more for an accurate review than for an inaccurate one. The quadratic form is to capture the assumption that the farther away from the revealed state, the faster is the disutility of deviation growing. The last term, $\alpha F(b)$, measures the utility from promoting the product, which comes from the firm's sponsorship. In this linear form, $F(b)$ is the proportion of customers whose type is below b, which is the demand for the product when the subscribers' belief is b. The

[5]Given a probability measure $\mathscr{P}$ over $\mathbb{R}$, $x \in \text{supp}(\mathscr{P})$ if for every open set U such that $x \in U$, $\mathscr{P}(u) > 0$.

linearity is assumed to capture the assumption that the blogger gets a fixed commission for each sale. In this paper, M and α are both exogenous. The constant C is assumed to be large enough so that the blogger's utility is higher with larger subscriber size. Moreover, the assumption spares us the need to check whether the blogger has an incentive to be a reviewer, and hence enables us to focus on analysing the blogger's incentive about the choice of messaging strategy.

Fix the prior λ and a subscriber belief mapping $b(\cdot;\lambda)$. The blogger chooses the messaging distribution $\Phi(s,\lambda)$ to maximize $E(u(b,\theta)|s,\lambda,\Phi(s,\lambda))$, the expected utility of the blogger conditional on the messaging distribution $\Phi(s,\lambda)$. The exact formula of the expected utility of the blogger when the belief mapping is b, the prior is λ, and is as below:

$$E(u(b(m;\lambda),\theta)|s,\lambda,\Phi(s,\lambda)) = q_s(\lambda)\int u(b(m;\lambda),1)d\Phi(\lambda,s)(m)+ [1-q_s(\lambda)]\int u(b(m;\lambda),0)d\Phi(\lambda,s)(m) \quad (3.2)$$

If $u(b,\theta)$ is concave in b, then $E(u(b,\theta)|s)$ is concave in b for every signal s. In this case, if the blogger can control the subscribers' belief, the belief b^* that maximizes the blogger's expected utility is characterized by the first-order condition,

$$b^* = q_s(\lambda) + \frac{\alpha}{2} f(b^*), \quad (3.3)$$

where $f(\cdot)$ is the probability density function of $F(\cdot)$ and $q_s(\lambda) = \Pr(\theta = 1|s,\lambda)$, the blogger's posterior belief about the product's quality upon receiving signal s. For simplicity, we assume the subscribers' types follow the standard uniform distribution. With this simplification, $f(b^*) = 1$, and $u(b,\theta)$ is concave in b. The second term on the right side is the belief markup the blogger would like to add if he could freely modulate his subscribers' belief. With F assumed to be the uniform distribution, the belief markup is constant. In general, the blogger is unable to achieve the optimal belief markup since the blogger's signal structure is common knowledge. Instead, the blogger chooses a message distribution that leads to a market belief that is optimal among the attainable beliefs.

3.3 Equilibrium analysis

Let $\Phi(s,\lambda) \in \Delta(\mathscr{M}(s))$ denote the strategy of the blogger when the realized signal is s and the prior is λ. Let $\Psi(m;\mu,\lambda) \in \{0,1\}$ denote the strategy of the subscriber with type μ when the realized message is m and the prior is λ. We make the following equilibrium definitions.

Definition 8 (Perfect Bayesian Equilibrium)**.** Fix a prior λ, a belief mapping $b(\cdot;\lambda)$, and a strategy profile $\left(\{\Phi(s;\lambda)\}_s, \{\Psi(\cdot;\mu)\}_\mu\right)$.

(a) The strategy profile is a **perfect Bayesian equilibrium** if the following three conditions hold.

Subscriber optimality For a subscriber of type μ and each message $m \in \mathscr{M}$, $\Psi(m;\mu) = \mathbf{1}(\mu \le b(m;\lambda))$.

Bayesian consistency If $m \in \operatorname{supp}(\Phi(\lambda))$, then $b(m;\lambda) = \Pr(\theta = 1|m, \{\Phi(s;\lambda)\}_s)$, i.e., the belief mapping agrees with the posterior belief about quality given a message and the blogger's strategy.[6]

Blogger optimality Given signal s, for almost every message $m \in \operatorname{supp}(\Phi(s;\lambda))$,

$$E(u(b(m;\lambda),\theta)|\lambda,s) = \sup_{\tilde{m}\in\mathscr{M}(s)} E(u(b(\tilde{m};\lambda),\theta)|\lambda,s).$$

(b) A perfect Bayesian equilibrium is a **pooling equilibrium** if $b(m) = \lambda$ with probability 1.

(c) A perfect Bayesian equilibrium is a **separating equilibrium** if $b(m) \in \{q_0(\lambda), q_1(\lambda)\}$ with probability 1.

For the sake of concision, we say an equilibrium instead of a perfect Bayesian equilibrium unless ambiguity arises. In equilibrium, we require the blogger's messaging strategy to maximize the blogger's expected payoff in each signal realization. For the subscribers, they purchase the reviewed product if and only if their belief about quality exceeds their types, and we require the beliefs given the blogger's message to be consistent with the blogger's strategy according to Bayes update rule.

We define a pooling equilibrium to be the case in which the subscribers do not get new information from the blogger whereas in a separating equilibrium, the subscribers would be able to know for sure whether the blogger's realized signal is high or low.

Before showing that the existence of a Perfect Bayesian equilibrium is guaranteed, we first look at the incentives of the blogger to take a sponsorship. It turns out that in this paper, the blogger always has a strictly incentive to accept the sponsorship.

Proposition 31 (Always sponsored)**.** *In every perfect Bayesian equilibrium, the blogger always has an incentive to accept the sponsorship.*

The intuition for the result is immediate. If in an equilibrium the blogger does not accept the sponsorship, the blogger can deviate by accepting sponsorship and keeping the message distribution unchanged. This deviation brings additional income to the blogger in each realized message and hence strictly improves the blogger's expected utility. Hence the blogger always has an incentive accept the sponsorship in equilibrium. The result assumes the existence of an equilibrium, which is validated by the following result.

Proposition 32 (Existence of pooling equilibrium)**.** *For every $\lambda \in (0,1)$, a pooling equilibrium exists.*

The result is consistent with the result from cheap talk that babbling, in which no information is communicated, is always an equilibrium outcome: once the belief mapping is fixed to the pooling belief, there is no way for the blogger to effectively change the subscribers' belief and hence no communication becomes an equilibrium. In general, there could be multiple equilibria as in

[6]The posterior probability is defined since we assume for every $m \in \operatorname{supp}(\Phi(\lambda))$, $\Pr(s = 1|m, \Phi(\cdot;\lambda))$ is defined, which makes $\Pr(\theta = 1|m, \Phi(\cdot;\lambda))$ defined.

cheap talk. However, we can show that whenever an equilibrium is not a pooling or separating equilibrium, then there exists a separating equilibrium in which the blogger's expected utility is weakly improved in both signal realizations.

Proposition 33 (Focus on separating and pooling)**.** *Fix a prior λ and a perfect Bayesian equilibrium. If the equilibrium is neither a pooling nor separating equilibrium, then there exists a separating equilibrium in which the blogger's expected utility is unchanged for signal $s = 0$ and strictly improved for signal $s = 1$.*

The proof has two steps. In the first step, we show that the blogger always has an incentive to choose the message that maximizes the subscribers' belief, provided that the belief mapping maps messages out of equilibrium path to the prior. The intuition is that with the sponsorship, the blogger has incentive to mark up the subscribers' belief from the blogger's belief, not to mark down, hence the incentive to maximize the subscribers' belief. The belief maximizing incentive of the blogger when the signal is high implies that the equilibrium belief mapping should be the same for almost every message the blogger sends in equilibrium when the belief is high. Since the belief mapping is Bayesian consistent with the blogger's strategy, if there exists a message $m \in \operatorname{supp}(\Phi(1;\lambda))$ but $m \notin \operatorname{supp}(\Phi(0;\lambda))$, for almost every such message, $b(m;\lambda) = \lambda$ because the probability of the signal being low given m would be 1. In this case, the blogger would be at least between the separating outcome and the partial pooling outcome. Since the blogger would always prefer a separating outcome when the signal is high, whenever an equilibrium is neither separating nor pooling, there exists a separating equilibrium that makes the blogger's payoff unchanged when the signal is low and strictly improved when the signal is high.

So far, Proposition 33 shows that for any equilibrium that is neither a pooling nor separating contract, there exists a separating equilibrium in which the blogger's expected utility is improved for each signal. Meanwhile, a separating outcome reveals all the available information about product quality to the subscribers. With these observations, this paper is to focus on pooling and separating equilibrium. Proposition 32 has shown that a pooling equilibrium always exists. The following result provides conditions under which a separating equilibrium exists. Specifically, if there is some prior that admits a separating equilibrium, then a separating equilibrium exists for a middle range of priors and only a pooling equilibrium exists when the priors are near 0 or 1.

Theorem 7 (Conditions for separation)**.** *If there exists some interior prior belief about quality admits a separating equilibrium, then there exists a separating equilibrium when the prior is $1/2$. Moreover, when a separating equilibrium exists for some prior, there exists $\underline{\lambda}$ and $\overline{\lambda}$ such that $0 \leq \underline{\lambda} \leq \overline{\lambda}$, $\underline{\lambda} + \overline{\lambda} = 1$, and a prior $\lambda \in (0,1)$ admits a separating equilibrium if and only if $\lambda \in \left[\underline{\lambda}, \overline{\lambda}\right]$.*

Since the blogger always prefers the separating outcome, it is only left to show in the proof that for $\lambda \in \left[\underline{\lambda}, \overline{\lambda}\right]$, when the signal is low, the blogger has no incentive to deviate in a strategy profile that leads to the separating outcome. Since by (3.2), for each signal, the blogger's expected utility is quadratic and concave in subscribers' belief b, when the signal is low, the blogger has no incentive to deviate in a separating strategy profile if $\Delta q(\lambda) := q_1(\lambda) - q_0(\lambda)$ is more than twice the optimal belief markup. The proof shows that $\Delta q(\lambda)$ is strictly increasing and then strictly decreasing for $\lambda \in (0,1)$. Since the optimal belief markup is constant for every interior prior thanks to the uniform distribution assumption, this intermediate result implies that there indeed exists $\underline{\lambda} \leq \overline{\lambda}$ such that the blogger has no incentive to deviate in a separating strategy profile if and only if $\lambda \in \left[\underline{\lambda}, \overline{\lambda}\right]$. In words, a separating equilibrium exists for mid-end product; for very low-end and high-end products, only a pooling equilibrium exists.

The proof of Theorem 7 also provides some very interpretable insights about the conditions for a separating equilibrium. When the blogger's signal is very informative in the sense that $\Delta q(s)$ is large, it becomes costly for the blogger to pretend to have received a high signal when the realized signal is actually low, because pretending to receive a high signal leads to a high probability of an inaccurate review. As a result, the blogger has an incentive to truthfully reveal a low signal if and only if the signal is very informative.

The result that $\underline{\lambda} + \overline{\lambda} = 1$ is immediate from the symmetric signal structure: since $\Delta q(\lambda) = \Delta q(1-\lambda)$, whenever a separating equilibrium exists for prior $\lambda \in (0,1)$, it also exists for prior $1-\lambda$.

Lastly, the claim that whenever a separating equilibrium exists for some interior prior, then it must exist when the prior is $1/2$ holds because it turns out that the blogger's signal is most informative when the prior is $1/2$. Indeed, the proof shows that $\Delta q(\lambda)$ is maximized at $\lambda = 1/2$. To see why the signal is most informative at $\lambda = 1/2$, consider the case where λ is near $1/2$ and near 0 or 1. When the prior is near $1/2$, the before-signal uncertainty is larger than when the prior is near 0 or 1. As a result, when a signal arrives, reduction in uncertainty is larger when the prior is near $1/2$ than when it is near 0 or 1. Indeed, for the prior very close to 0 or 1, the change to the belief about quality is very small. Therefore, a signal is more informative when the prior is close to the half than when it is near the two ends. Consequently, whenever a separating equilibrium exists for some $\lambda \neq 1/2$, such an equilibrium exists for products whose priors are in the medium range. For products that are perceived to be very good or very bad before the blogger's signal, the blogger has an incentive to pretend to have received a high signal because the blogger's signal does not change the belief about quality very much.

Since a pooling equilibrium always exists, it is of interest to understand how the blogger's expected utility compare between a pooling and a separating equilibrium. Since by the proof of Proposition 33, the blogger would also wishes to maximize the Bayesian-consistent belief about quality when the signal of high, the expected utility comparison will mostly focus on the analysis for the case where the signal is low. In other words, we wish to know which between a pooling and a separating equilibrium the blogger would choose when the signal is low and a separating equilibrium exists.

Since the optimal markup to the blogger is positive, intuitively, when the signal is low, the blogger would prefer a separating equilibrium to a pooling equilibrium if the pooling belief is sufficiently different from the low-signal belief: when the two beliefs are very different, it is costly for the blogger to lead the subscribers to a belief much higher than the blogger's belief because of the high probability of an inaccurate review. The following result shows that the blogger prefers the separating equilibrium to a pooling equilibrium for some middle range of quality priors.[7]

Theorem 8 (Condition for separation preference)**.** *Assume some interior prior admits a separating*

[7] See Ying Chen, Kartik, and Sobel (2008) for a more in-depth analysis of equilibria selection in cheap-talk games.

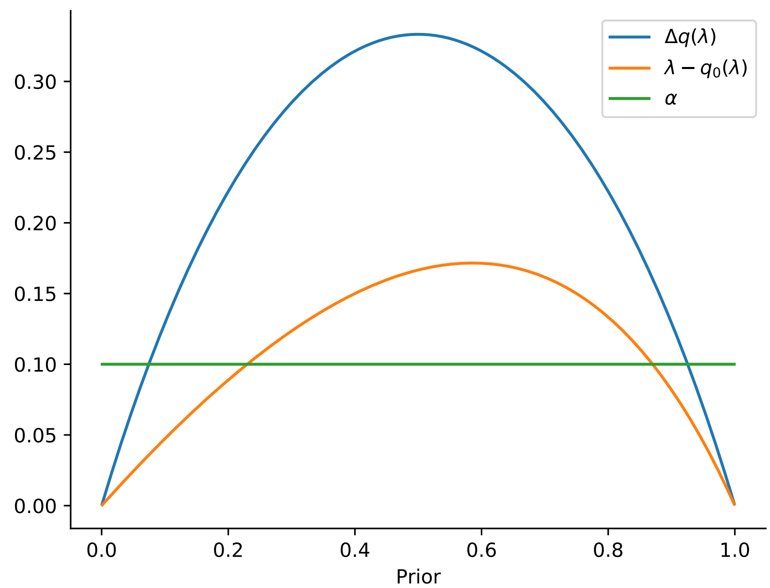

Figure 3.2: Conditions for separating equilibrium and for preference for separating equilibrium

equilibrium and let $\underline{\lambda}$ and $\overline{\lambda}$ be as defined in Theorem 7. Either the blogger always strictly prefers a pooling equilibrium to a separating equilibrium for every $\lambda \in (0,1)$, or there exists $\underline{\lambda} \leq \underline{l} \leq \overline{l} \leq \overline{\lambda}$ such that for $\lambda \in (0,1)$, when the signal is low, the blogger prefers a separating equilibrium to a pooling equilibrium if and only if $\lambda \in \left[\underline{l}, \overline{l}\right]$.

To show whether a blogger would prefer a separating equilibrium to a pooling equilibrium is equivalent to showing whether the blogger would have an incentive to choose a message that the subscribers' belief mapping maps to the pooling belief. With the similar reasoning to that of Theorem 7, the blogger would have such an incentive when the signal is low when λ is sufficiently above $q_0(\lambda)$ so that it would be costly for the blogger to choose the pooling belief. Then, similar to the proof of Theorem 7, the proof of Theorem 8 shows that $\lambda - q_0(\lambda)$ is decreasing and then increasing in $\lambda \in (0,1)$.

Figure 3.2 draws $\Delta q(\lambda 0)$ and $\lambda - q_0(\lambda)$, with $r = 2/3$ and $\alpha = 0.1$. The horizontal line in the graph is twice the optimal markup. The graph illustrates the result that a separating equilibrium exists for a middle range of quality priors and the blogger prefers a separating equilibrium to a pooling equilibrium for a smaller middle range of quality priors. For the smaller middle range of priors, when the signal is low, the blogger would have an incentive to truthfully reveal the realized signal even if there were a message that would lead the subscribers to have the pooling belief. Another observation from the graph is that whereas the curve of $\Delta q(\lambda)$ centres around $1/2$, the curve of $\lambda - q_0(\lambda)$ leans more above $1/2$. The intuition for the latter's inclination above $1/2$ is that when $\lambda > 1/2$, the prior is already in favour of a high quality. In this case, a low signal would come as a surprise, which would lead to a larger change from the prior than a high signal. Indeed, the next result formally shows that the exact middle between $\underline{l}$ and $\overline{l}$ is always above $1/2$.

Proposition 34 (More preference for separation when high prior)**.** *Assume there exist $\underline{l}$ and $\bar{l}$ as defined in Theorem 8. If $\underline{l} < \bar{l}$, then $\bar{l} > 1/2$ and $\underline{l} + \bar{l} > 1$.*

Here are some additional statistical intuitions of the result. When $\lambda < 1/2$, since a low signal is more likely and $\lambda = \Pr(s = 1|\lambda)q_1(\lambda) + \Pr(s = 0|\lambda)q_0(\lambda)$, $\lambda - q_0(\lambda) < q_1(\lambda)_\lambda$. By symmetry, $(1 - \lambda) - q_0(1 - \lambda) = q_1(\lambda) - \lambda$. Therefore, if $\lambda - q_0(\lambda)$ is large enough for the blogger to prefer a separating equilibrium to a pooling equilibrium with low signal when the prior is $\lambda < 1/2$, then the same is strictly true for the blogger when the prior is $1 - \lambda$, which leads to the conclusion that $\underline{l} + \bar{l} > 1$.

We now provide some comparative statics results regarding the conditions for the existence of a separating equilibrium. The following result shows that, fixing everything else constant, higher signal precision and lower commission rate widen the middle range of priors that admit a separating equilibrium.

Proposition 35 (Comparative statics)**.** *Fixing everything else constant, $\underline{l}$ and $\underline{\lambda}$ as defined in Theorem 7 and Theorem 8 are increasing decreasing in r and increasing in α; $\bar{l}$ and $\bar{\lambda}$ are increasing in r and decreasing in α.*

The proof of this result is very intuitive. When the signal precision improves, signals become more informative in the sense that change in belief from the prior is larger. As a result, it becomes more costly for the blogger to pretend to have received a high signal and to choose a message that leads to the pooling belief. Therefore, higher r makes it easier for a separating equilibrium to exist and for the blogger to prefer a separating equilibrium to a pooling equilibrium when the signal is low.

When α, the sales commission rate, gets larger, the blogger has a larger optimal belief markup, which makes the requirement on signal informativeness more demanding. As a result, higher α makes it harder for a separating equilibrium to exist and for the blogger to prefer a separating equilibrium to a pooling equilibrium when the signal is low.

The results so far concern the incentives of the blogger conditional on signal realization. Given the prior, it is of interest to know what equilibrium the blogger would choose if the blogger could choose an equilibrium before the signal arrives. The following result shows that the blogger has a commitment issue: before the signal arrival, the blogger always prefers a separating equilibrium if there one exists, but Theorem 8 shows that for some quality priors, a separating equilibrium exists but the blogger would prefer a pooling equilibrium when the signal is low.

Proposition 36 (Commitment issue)**.** *Fix a prior that admits a separating equilibrium. Before the signal arrives, the blogger is always better off with a separating equilibrium than with a pooling equilibrium.*

Proof. By (3.1) and the distributional assumption of $F(\cdot)$, $\alpha F(b)$ is linear in the subscribers' belief b. Since in an equilibrium where the belief mapping is $b(\cdot; \lambda)$ and the blogger's strategy is $\Phi(\cdot; \lambda)$, the belief mapping is Bayesian consistent, we have

$$E\left[E(b(m; \lambda)|\Phi(s; \lambda), \lambda)\right] = \lambda. \tag{3.4}$$

As the expected subscribers' belief is unchanged in every equilibrium for the same fixed prior,

$$E\left[\alpha E(F(b(m;\lambda))|\Phi(s;\lambda),\lambda)\right] = \alpha\lambda. \tag{3.5}$$

For the loss minimization of deviation from the revealed state in (3.2), note that in a separating equilibrium,

$$q_s(\lambda) = \arg\inf_{b\in(0,1)} E\left[(b-\theta)^2|s,\lambda\right]. \tag{3.6}$$

Since the loss function in the expectation operator is strictly convex, the solution to the minimization problem is unique, and hence

$$E\left[(q_s(\lambda)-\theta)^2|s,\lambda\right] < E\left[(\lambda-\theta)^2|s,\lambda\right], \tag{3.7}$$

for each signal s. Therefore, in a separating equilibrium, the expected loss is strictly lower than that in a pooling equilibrium. Hence the blogger always prefers a separating equilibrium to a pooling equilibrium before the signal realization. □

Thanks to the restriction that the subscribers' belief mapping in an equilibrium must be Bayesian consistent, the expected subscribers' belief before the signal realization in every equilibrium is the prior. Since the blogger's expected utility from sales commissions is linear with respect to subscribers' belief, the blogger's expected utility from sales commissions is the same in every equilibrium. With regard to the loss of deviating from the revealed state as shown in (3.2), the separating equilibrium reaches the unique minimal expected loss conditional signal realization. Therefore, to the blogger before the signal arrives, a separating equilibrium is able to deliver a strictly better loss minimization than does a pooling equilibrium. Hence to the blogger before the signal realization, a separating equilibrium is always preferred to a pooling equilibrium when a separating equilibrium exists for the prior of the reviewed product.

3.4 Conclusion

This paper has studied the reviewing incentive of a blogger that sends a review message to the blogger's subscribers, with the reviewed product's firm offering to sponsor the blogger's review through sales commissions. The paper shows that the blogger's incentive to truthfully communicate the private signal to the subscribers is closely linked to the informativeness of the blogger's signal: when the signal is more informative, then the blogger is more likely to provide an honest review; when the signal is less informative, then the blogger is more likely to pretend to receive a high signal even if the realized the signal is low. When the blogger's signal structure is binary and symmetric, it has been shown that with a fixed signal precision, the blogger's signal is informative enough for truthful communication when the prior belief about quality is within some middle range of priors. In other words, the blogger has an incentive for an honest review when the product is a mid-end product and for a biased review when the product is very low-end or high-end.

The paper does have some limitations. For example, the paper uses a very simplified signal structure that consists of one binary signal about the product's overall quality. An alternative approach is to treat the review process as receiving a number of (possibly costly) binary signals about different attributes of the product. There have been a number of studies from consumer search, such as Branco, Sun, and Villas-Boas (2012) and Weitzman (1979), that look at the optimal sequential information acquisition problem of an agent. In the context of product reviewing, if a product has multiple attributes and each attributes require a costly signal to learn more about that attribute, then the blogger would need to solve an optimal stopping problem as in other consumer search models. After the information acquisition stage, the blogger needs to work out a message for the subscribers and then the subscribers are going to form a belief about the product. In this case, the blogger is in a multidimensional cheap-talk game and the blogger's information acquisition strategy is likely to be affected by the blogger's communication strategy. We leave this type of more complex yet important analysis for future research.

3.A Proofs

3.A.1 Proof of Proposition 31

Proof. Suppose there is an equilibrium in which a blogger decides not to accept the sponsorship after receiving the signal. If the blogger deviates by accepting the sponsorship and chooses the same message distribution as the one the blogger chooses in equilibrium, the distribution of the subscribers' beliefs is unchanged. Now for each realized message, the blogger gets additional income from the sales commissions, making the blogger strictly better off. Therefore, it is profitable for the blogger to deviate to accept the sponsorship, contradicting the definition of the equilibrium. Thus in every equilibrium, the blogger always accepts the sponsorship. □

3.A.2 Proof of Proposition 32

Proof. Consider the strategy profile where $\Phi(0;\lambda)(m) = \Phi(1;\lambda)(m)$ for every $m \in \mathscr{M}(1)$ and the belief mapping is Bayesian consistent with the blogger's strategy. If $m \notin \text{supp}(\Phi(0;\lambda)) \cup \text{supp}(\Phi(1;\lambda))$, let $b(m;\lambda) = 0$. For each signal realization, the blogger's optimal subscriber belief is above the blogger's belief. Moreover, as the blogger's expected utility is concave and quadratic in the subscribers' belief, having the zero belief is never optimal. Therefore, for each signal realization, the blogger has no incentive to choose a message not in the support of the proposed strategy. Since the belief mapping is constant for messages in the support of the blogger's messaging distributions, the proposed strategy is trivially optimal to the blogger for both signal realizations. Therefore, a pooling equilibrium always exists. □

3.A.3 Proof of Proposition 33

Proof. We can show that for given the subscribers' belief about quality conditional on realized message, it is optimal for the blogger choose a message that maximizes the subscribers' belief, which is summarized in the following lemma.

Lemma 16 (Optimality of belief maximization with high signal)**.** *Fix a prior λ and strategy profile where the blogger's strategy is $\Phi(\cdot;\lambda)$. Let $b(\cdot;\lambda)$ be a belief mapping that is Bayesian consistent with the blogger's strategy and $b(m;\lambda) = \lambda$ for every $m \notin supp(\Phi(\lambda))$. When $s = 1$, sending any message from* $\arg\sup_{m \in \mathcal{M}(1)} b(m;\lambda)$ *is optimal to the blogger.*

Proof of lemma. Fix a prior λ. By (3.3) and Proposition 31, when the signal is high, the optimal subscriber belief to the blogger is strictly above $q_1(\lambda)$. Moreover, by (3.2), the blogger's expected utility is concave and quadratic in the subscriber's belief. Therefore, $E(u(q_1(\lambda),\theta)|s = 1,\lambda) > E(u(b,\theta)|s = 1,\lambda)$ for every $b < q_1(\lambda)$. Thus it is optimal for the blogger to assign measure zero to the complement of $\text{supp}(\Phi(\lambda))$ when the signal is high, i.e., the blogger has no incentive to send a message not in $\text{supp}(\Phi(\lambda))$ when the signal is high.

Since the belief mapping $b(\cdot;\lambda)$ is Bayesian consistent with the blogger's strategy, for $m \in \text{supp}(\Phi(\lambda))$, $b(m;\lambda) \in [q(0;\lambda), q(1;\lambda)]$, and thus

$$E(u(b(m;\lambda),\theta)|\lambda, s, \Phi(s,\lambda)) \leq E(u(b(m^*;\lambda),\theta)|\lambda, s, \delta_{m^*}),$$

where $m^* = \arg\sup_{m^* \in \mathcal{M}(s)} b(m;\lambda)$ and δ_{m^*} is the Dirac measure at m^*.[8] Therefore, it is optimal for the blogger to choose the message that maximizes the subscribers' belief when the signal is high and the belief mapping is Bayesian consistent. The proof of the lemma is complete. □

By Lemma 16, in an equilibrium, for almost every m_1 and m_2 in $\text{supp}(\Phi(1;\lambda))$ with respect to $\Phi(1;\lambda)$, $b(m_1;\lambda) = b(m_2;\lambda)$, i.e., the belief mapping is constant among messages that the blogger finds it optimal to send when $s = 1$. If for almost every $m \in \text{supp}(\Phi(1;\lambda))$, $b(m;\lambda) = \lambda$, then the equilibrium is necessarily a pooling equilibrium since $\Phi(0;\lambda)$ must integrate to 1 on $\text{supp}(\Phi(1;\lambda))$ for $b(m;\lambda) = \lambda$.

Now assume $b(m;\lambda) \neq \lambda$ for almost every $m \in \text{supp}(\Phi(1;\lambda))$ with respect to $\Phi(1,\lambda)$. In this case, it is necessary that the event $b \in \text{supp}(\Phi(0,\lambda))$ and $b(m;\lambda) = q_0(\lambda)$ has a strictly positive measure with respect to $\Phi(0,\lambda)$. If the measure of the event is 1 with respect to $\Phi(0;\lambda)$, then the equilibrium is a pooling equilibrium. If the measure is strictly less than one, then $\text{supp}(\Phi(0;\lambda)) - \text{supp}(\Phi(1;\lambda))$ has a strictly positive measure w.r.t $\Phi(0;\lambda)$. By the blogger's optimality condition, when $s = 0$, the blogger is indifferent between $\Phi(0;\lambda)$ restricted to $\text{supp}(\Phi(0;\lambda)) \setminus \text{supp}(\Phi(1;\lambda))$ and the same measure restricted to $\text{supp}(\Phi(1;\lambda))$. In the first restriction, when $s = 0$, the blogger is getting the expected payoff in a separating equilibrium if there exists one. Moreover, since $q_1(\lambda) > b(m;\lambda)$ for a.e. $m \in \text{supp}(\Phi(1;\lambda))$, and since $E(u(b;\theta)|\lambda, s)$ is concave and quadratic in b, for a.e. $m \in \text{supp}(\Phi(1;\lambda))$,

$$E(u(q_0(\lambda),\theta)|\lambda, s = 0) = E(u(b(m;\lambda),\theta)|\lambda, s = 0) > E(u(q_1(\lambda),\theta)|\lambda, s = 0), \tag{3.8}$$

which implies that the blogger will have no incentive to deviate in a strategy profile that would be a separating equilibrium if it is a PBE. Since the blogger would always prefer a separating equilibrium when $s = 1$, the blogger has no incentive to deviate in such a strategy profile, either. Additionally, since in such a strategy profile, $b(m;\lambda) = q_1(\lambda)$ for a.e. $m \in \text{supp}(\Phi(1;\lambda))$, the blogger's expected utility is strictly improved in this strategy profile. Therefore, when an equilibrium is neither a pooling nor separating equilibrium, there exists a separating equilibrium in which the blogger's expected utility is unchanged when $s = 0$ and strictly improved when $s = 1$. □

[8]Formally, δ_{m^*} is a probability measure such that for every measurable set U, $\delta_{m^*}(U) = \mathbf{1}(m^* \in U)$.

3.A.4 Proof of Theorem 7

Proof. Fix a prior λ, a strategy profile, and a belief mapping $b(\cdot;\lambda)$ such that subscriber optimal and Bayesian consistency hold. Assume the strategy profile leads to the separating outcome and let $\Phi(\cdot;\lambda)$ be the blogger strategy in this profile. For $m \notin \text{supp}(\Phi(\lambda))$, assume $b(m;\lambda) = q_0(\lambda)$.

As the subscriber type distribution is assumed to be the standard uniform distribution, $F(x) = x$ and $f(x) = 1$ for $x \in [0,1]$. By (3.3), when the prior is λ, for each signal, the blogger's optimal subscriber belief is $\min\{q_s(\lambda) + \frac{\alpha}{2}, 1\}$. Since by (3.2), the blogger's expected utility is quadratic and concave, when the signal is low, the blogger has no incentive to deviate to a message m where $b(m;\lambda) = q_1(\lambda)$ if and only if $q_1(\lambda) - q_0(\lambda) \geq \alpha$. By the Bayes rule,

$$\Delta q(\lambda) := q_1(\lambda) - q_0(\lambda) = \frac{\lambda(1-\lambda)(2r-1)}{[r+\lambda(1-2r)][1-r+\lambda(2r-1)]}, \tag{3.9}$$

and the difference's first-order derivative with respect to λ is

$$\frac{\partial \Delta q(\lambda)}{\partial \lambda} = \frac{r(1-r)(2r-1)(1-2\lambda)}{[r+\lambda(1-2r)]^2[1-r+\lambda(2r-1)]^2}. \tag{3.10}$$

Note that the numerator is always non-negative. Thus to solve $\alpha q(\lambda)/\partial\lambda \geq 0$, we have

$$\frac{\partial \Delta q(\lambda)}{\partial \lambda} \geq 0 \Longrightarrow 0 \leq \lambda \leq \frac{1}{2}. \tag{3.11}$$

Therefore, $\Delta q(\lambda)$ is increasing for $\lambda \in (0, 1/2)$ and decreasing for $\lambda \in (1/2, 1)$. If $\Delta q(1/2) < \alpha$, then for every $\lambda \in (0,1)$, there is no separating equilibrium. Otherwise, there exists $0 \leq \underline{\lambda} < 1/2 < \overline{\lambda} \leq 1$ such that $\Delta q(\lambda) \geq \alpha$ if and only if $\lambda \in \left[\underline{\lambda}, \overline{\lambda}\right]$. Therefore, a separating equilibrium exists if and only if $\lambda \in \left[\underline{\lambda}, \overline{\lambda}\right]$.

To see that $\underline{\lambda} + \overline{\lambda} = 1$, it suffices to show that if a prior $\lambda \in (0,1)$ admits a separating equilibrium, then the prior $1-\lambda$ also admits a separating equilibrium. To see this, note that $\Delta q(\lambda) = \Delta q(1-\lambda)$. Therefore, if $\Delta q(\lambda) \geq \alpha$, then $\Delta q(1-\lambda) \geq \alpha$, meaning that when the prior is $1-\lambda$, there exists a separating equilibrium. Thus $\underline{\lambda} + \overline{\lambda} = 1$. □

3.A.5 Proof of Theorem 8

Proof. Pick a prior λ that admits a separating equilibrium. When the signal is low, to compare whether the blogger would prefer a separating equilibrium to a pooling equilibrium is the same to see whether the blogger has an incentive to send a message which the belief mapping maps to λ. With the same reasoning in the proof of Theorem 7, when the signal if low, the blogger prefers $q_0(\lambda)$ to λ if and only if $\lambda - q_0(\lambda) \geq \alpha$. To prove this, we are to show that $\lambda - q_0(\lambda)$ is strictly increasing and then decreasing for $\lambda \in (0,1)$. To show the pattern, first we have

$$\lambda - q_0(\lambda) = \lambda - \frac{(1-r)\lambda}{\lambda(1-r)+(1-\lambda)r} = \frac{\lambda(1-\lambda)(2r-1)}{r+\lambda(1-2r)}.$$

If we take the first-order derivative of $\lambda - q_0(\lambda)$, we get

$$\frac{\partial}{\partial\lambda}\left[\lambda - q_0(\lambda)\right] = \frac{(2r-1)([(2r-1)\lambda^2 - 2r\lambda + r])}{[r+\lambda(1-2r)]^2}.$$

The denominator is non-negative, and it can be shown that the numerator is strictly positive if and only if $\lambda < \frac{r-\sqrt{r(1-r)}}{2r-1}$. Denote the right-hand of the inequality by l^*. If $\lambda - q_0(l^*) < \alpha$, then when the signal is low, the blogger always strictly prefers a pooling equilibrium to a separating equilibrium for each interior prior. Otherwise, we can find $\underline{l} < \bar{l}$ such that $\lambda - q_0(\lambda) \geq \alpha$ if and only if $\lambda \in \left[\underline{l}, \bar{l}\right]$, and hence the blogger would prefer a separating equilibrium to a pooling equilibrium when the signal is low if the prior is in that range. □

3.A.6 Proof of Proposition 34

Proof. Assume $\underline{l}$ and $\bar{l}$ as defined in Theorem 8 exist. Note that $\Pr(s=1|\lambda) \geq 1/2$ if and only if $\lambda \geq 1/2$. Therefore, since $\lambda = \Pr(s=1|\lambda)q_1(\lambda) + \Pr(s=0|\lambda)q_0(\lambda)$, $q_1(\lambda) - \lambda \leq q_0(\lambda)$ if and only if $\lambda \geq 1/2$.

Assume $\underline{l} < \bar{l}$. If $\underline{l} \geq 1/2$, then $\bar{l} > 1/2$ and clearly $\underline{l} + \bar{l} > 1$. Assume instead that $\underline{l} < 1/2$ and pick a prior $l \in \left[\underline{l}, \bar{l}\right]$ such that $l < 1/2$. Note that for every $\tilde{\lambda} \in (0,1)$,

$$q_1(\tilde{\lambda}) = 1 - q_0(1-\tilde{\lambda}) \tag{3.12}$$

which implies $\Delta q(\tilde{\lambda}) = \Delta q(1-\tilde{\lambda})$. Therefore, if $\Delta q(\lambda) \geq \alpha$, then $\Delta q(1-\lambda) \geq \alpha$. In other words, when the prior is $1-\lambda$, a separating equilibrium exists. Moreover, by (3.12),

$$\lambda - q_0(\lambda) = q_1(1-\lambda) - (1-\lambda). \tag{3.13}$$

If $\lambda - q_0(\lambda) \geq \alpha$, then since $1-\lambda > 1/2$,

$$(1-\lambda) - q_0(1-\lambda) > q_1(1-\lambda) - (1-\lambda) = \lambda - q_0(\lambda) \geq \alpha. \tag{3.14}$$

Therefore, when the prior is $1-\lambda$, a separating equilibrium exists and the blogger prefers a separating equilibrium to a pooling equilibrium. This implies $\bar{l} \geq 1/2$ and $\underline{l} + \bar{l} \geq 1$. □

3.A.7 Proof of Proposition 35

Proof. Fix $\lambda \in (0,1)$. Since $q_0(\lambda)$ is decreasing in r and $q_1(\lambda)$ is increasing in r, both $\Delta q(\lambda)$ and $\lambda - q_0(\lambda)$ are increasing in r. By the proof of Theorem 7 and 8, $\underline{\lambda}$ and $\underline{l}$ are decreasing in r, whereas $\bar{l}$ and $\bar{\lambda}$ are increasing in r.

When α gets larger, the blogger's optimal belief markup gets larger, which raises the requirement on $\Delta q(\lambda)$ larger for a separating equilibrium to exists. Similarly, larger α also raises the requirement on $\lambda - q_0(\lambda)$ for the blogger to prefer a separating equilibrium when the signal is low. Therefore, $\underline{\lambda}$ and $\underline{l}$ are increasing in α, whereas $\bar{l}$ and $\bar{\lambda}$ are decreasing in α. □

Bibliography

Adiri, Igal and Ury Yechiali (1974). “Optimal priority-purchasing and pricing decisions in non-monopoly and monopoly queues”. In: *Operations Research* 22.5, pp. 1051–1066.

Alperstein, Hanna (1988). “Note—optimal pricing policy for the service facility offering a set of priority prices”. In: *Management Science* 34.5, pp. 666–671.

Balachandran, K. R. (1972). “Purchasing priorities in queues”. In: *Management Science* 18.5-part-1, pp. 319–326.

Balmaceda, Felipe et al. (2016). “Bounds on the welfare loss from moral hazard with limited liability”. In: *Games and Economic Behavior* 95, pp. 137–155.

Basu, Amiya K et al. (1985). “Salesforce compensation plans: An agency theoretic perspective”. In: *Marketing Science* 4.4, pp. 267–291.

Bergstrom, Theodore, Lawrence Blume, and Hal Varian (1986). “On the private provision of public goods”. In: *Journal of Public Economics* 29.1, pp. 25–49.

Biyalogorsky, Eyal, Eitan Gerstner, and Barak Libai (2001). “Customer referral management: Optimal reward programs”. In: *Marketing Science* 20.1, pp. 82–95.

Bloch, Francis and Armando Gomes (2006). “Contracting with externalities and outside options”. In: *Journal of Economic Theory* 127.1, pp. 172–201.

Bolton, Patrick and Mathias Dewatripont (2005). *Contract theory*. MIT press.

Branco, Fernando, Monic Sun, and J Miguel Villas-Boas (2012). “Optimal search for product information”. In: *Management Science* 58.11, pp. 2037–2056.

Carroll, Gabriel (2012). “When are local incentive constraints sufficient?” In: *Econometrica* 80.2, pp. 661–686.

Chen, Ying, Navin Kartik, and Joel Sobel (2008). “Selecting cheap-talk equilibria”. In: *Econometrica* 76.1, pp. 117–136.

Chen, Yubo and Jinhong Xie (2005). “Third-party product review and firm marketing strategy”. In: *Marketing science* 24.2, pp. 218–240.

Crawford, Vincent P and Joel Sobel (1982). “Strategic information transmission”. In: *Econometrica*, pp. 1431–1451.

Csorba, Gergely (2008). “Screening contracts in the presence of positive network effects”. In: *International Journal of Industrial Organization* 26.1, pp. 213–226.

Farrell, Joseph and Matthew Rabin (1996). “Cheap talk”. In: *Journal of Economic perspectives* 10.3, pp. 103–118.

Garnefeld, Ina et al. (2013). “Growing existing customers’ revenue streams through customer referral programs”. In: *Journal of Marketing* 77.4, pp. 17–32.

Gomes, Armando (2005). “Multilateral contracting with externalities”. In: *Econometrica.*
Grossman, Sanford J and Oliver D Hart (1983). “An Analysis of the Principal-Agent Problem”. In: *Econometrica* 51.1, pp. 7–46.
Guesnerie, Roger and Jean-Jacques Laffont (1984). “A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm”. In: *Journal of Public Economics* 25.3, pp. 329–369.
Hassin, Refael and Moshe Haviv (1997). “Equilibrium threshold strategies: The case of queues with priorities”. In: *Operations Research* 45.6, pp. 966–973.
Hölmstrom, Bengt (1979). “Moral hazard and observability”. In: *The Bell Journal of Economics*, pp. 74–91.
Hwang, Yoori and Se-Hoon Jeong (2016). ““This is a sponsored blog post, but all opinions are my own”: The effects of sponsorship disclosure on responses to sponsored blog posts”. In: *Computers in Human Behavior* 62, pp. 528–535.
Innes, Robert D (1990). “Limited liability and incentive contracting with ex-ante action choices”. In: *Journal of Economic Theory* 52.1, pp. 45–67.
Jewitt, Ian (1988). “Justifying the first-order approach to principal-agent problems”. In: *Econometrica*, pp. 1177–1190.
Jin, Liyin and Yunhui Huang (2014). “When giving money does not work: the differential effects of monetary versus in-kind rewards in referral reward programs”. In: *International Journal of Research in Marketing* 31.1, pp. 107–116.
Kamada, Yuichiro and Aniko Öry (2020). “Contracting with word-of-mouth management”. In: *Management Science.*
Katz, Michael L. and Carl Shapiro (1986). “Technology adoption in the presence of network externalities”. In: *Journal of Political Economy* 94.4, pp. 822–841.
Kim, Son Ku (1997). “Limited liability and bonus contracts”. In: *Journal of Economics & Management Strategy* 6.4, pp. 899–913.
Kuksov, Dmitri and J Miguel Villas-Boas (2010). “When more alternatives lead to less choice”. In: *Marketing Science* 29.3, pp. 507–524.
Kumar, Viswanathan, J Andrew Petersen, and Robert P Leone (2007). “How valuable is word of mouth?” In: *Harvard Business Review* 85.10, p. 139.
Laffont, Jean-Jacques and Jean Tirole (1986). “Using cost observation to regulate firms”. In: *Journal of Political Economy* 94.3, Part 1, pp. 614–641.
Leduc, Matt V, Matthew O Jackson, and Ramesh Johari (2017). “Pricing and referrals in diffusion on networks”. In: *Games and Economic Behavior* 104, pp. 568–594.
Lobel, Ilan, Evan Sadler, and Lav R Varshney (2017). “Customer referral incentives and social media”. In: *Management Science* 63.10, pp. 3514–3529.
Maskin, Eric and John Riley (1984). “Monopoly with incomplete information”. In: *The RAND Journal of Economics* 15.2, pp. 171–196.
McAfee, R Preston and John McMillan (1987). “Competition for agency contracts”. In: *The RAND Journal of Economics*, pp. 296–307.
Melumad, Nahum D and Stefan Reichelstein (1989). “Value of communication in agencies”. In: *Journal of Economic Theory* 47.2, pp. 334–368.

Mirrlees, James A (1976). “The optimal structure of incentives and authority within an organization”. In: *The Bell Journal of Economics*, pp. 105–131.

— (1999). “The theory of moral hazard and unobservable behaviour: Part I”. In: *The Review of Economic Studies* 66.1, pp. 3–21.

Park, Jeong-Yeol and SooCheong Shawn Jang (2013). “Confused by too many choices? Choice overload in tourism”. In: *Tourism Management* 35, pp. 1–12.

Poblete, Joaquin and Daniel Spulber (2012). “The form of incentive contracts: agency with moral hazard, risk neutrality, and limited liability”. In: *The RAND Journal of Economics* 43.2, pp. 215–234.

Raju, Jagmohan S and V Srinivasan (1996). “Quota-based compensation plans for multiterritory heterogeneous salesforces”. In: *Management Science* 42.10, pp. 1454–1462.

Rao, Ram C (1990). “Compensating heterogeneous salesforces: some explicit solutions”. In: *Marketing Science* 9.4, pp. 319–341.

Rasmusen, Eric B., J Mark Ramseyer, and John S Wiley Jr (1991). “Naked exclusion”. In: *The American Economic Review*, pp. 1137–1145.

Rochet, Jean-Charles (1987). “A necessary and sufficient condition for rationalizability in a quasi-linear context”. In: *Journal of Mathematical Economics* 16.2, pp. 191–200.

Rogerson, William P (1985). “The first-order approach to principal-agent problems”. In: *Econometrica*, pp. 1357–1367.

Schmitt, Philipp, Bernd Skiera, and Christophe Van den Bulte (2011). “Referral programs and customer value”. In: *Journal of Marketing* 75.1, pp. 46–59.

Segal, Ilya (1999). “Contracting with externalities”. In: *The Quarterly Journal of Economics* 114.2, pp. 337–388.

Segal, Ilya and Michael D. Whinston (2003). “Robust predictions for bilateral contracting with externalities”. In: *Econometrica* 71.3, pp. 757–791.

Shi, Zijun, Kaifu Zhang, and Kannan Srinivasan (2019). “Freemium as an optimal strategy for market dominant firms”. In: *Marketing Science* 38.1, pp. 150–169.

Weitzman, Martin L (1979). “Optimal search for the best alternative”. In: *Econometrica*, pp. 641–654.