TEL AVIV UNIVERSITY

Notes on Rational Queueing

by

Ran I. Snitkovsky

Under supervision of **Prof. Refael Hassin**

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the

Raymond & Beverly Sackler Faculty of Exact Sciences
the Department of Statistics and Operations Research

August 28, 2019

Abstract

This thesis deals with models of strategic behavior of customers in service systems (i.e., *rational queueing*). It touches on two complementary parts in the rational queueing literature: *observable* and *unobservable* queueing models. In the context of observable queues, we point out properties that imply the infamous Naor's Inequality, suggest model applications for our findings, and further provide simple examples where the inequality does not hold. In the context of unobservable queues, we investigate two different concrete models: In one model we introduce a noncooperative multi-player game of individual rational users sending data-packets in a Cognitive Radio Network with the opportunity of spectrum sensing. In the other model we study the impact of tipping in a service facility on the server's tipping wage, in the presence of an endogenously formed social norm. We provide a comprehensive Game Theoretic and Queueing analysis in both models and discuss equilibrium and socially optimal behavior.

Acknowledgements

I wish to express my deep gratitude to my adviser, Refael Hassin (Rafi), for all the years of teaching, guidance, and support. Rafi was always available for advice and comments on my work, with great patience and care. Rafi expressed much thoughtfulness, allowed me freedom and encouragement to work on my own ideas, still never compromising on academic standards, or sparing criticism when it is due.

I am very grateful to Laurens G. Debo for hosting me at Tuck School of Business at Dartmouth College. During the collaboration on our project Laurens became an unofficial mentor for me.

Acknowledgement is also due to all of my great teachers, and especially to Moshe Haviv and Uri Yechiali for their advice and guidance; to my friends – my fellow-student Michal Ben-Elli and my colleagues Liron Ravner and Yoav Kerner for numerous fruitful discussions.

I am grateful for the financial support I have received over the last few years from the School of Mathematical Sciences at Tel Aviv University, the Operations Research Society of Israel, the UJA Federation Selim and Rachel Benin Scholarship Fund, the Association of European OR Societies and to the Israel Science Foundation (Grant No. 355/15).

Last but not least, I thank my family – my parents, in particular, for their love and support throughout my studies; and ultimately, my biggest gratitude goes to my beloved wife Noa, without whom I would have never embarked upon this journey.

Contents

Chapter 1

Introduction

Coping with congestion is a task we are faced with on a daily basis, in traffic, in healthcare, at the grocery store, or even at our office kitchenette. Congestion affects us even when its presence is intangible, for instance while surfing the internet, waiting for an Uber driver, etc. Motives for avoiding congestion include saving time and reducing expenses, and, as much as for us, the customers, congestion is a far-reaching phenomenon of concern for system operators and managers.

In a typical metropolitan morning, every single commuter among millions chooses, out of a large set of alternatives, when and how to get to work. The commute-time of each such traveler depends on the over-all congestion brought about in the city's transportation infrastructure, which is the outcome of the commuters' preferences. Is the infrastructure utilized optimally? And how would the commute be affected by changes in the infrastructure, for example, like paving a new road? These questions call for devising a suitable mathematical model. The focus of my research is on the modeling and analysis of such systems, professionally known as Queueing Systems. Specifically, it revolves around interactions of decision-making customers and/or service providers, and the way they induce congestion in queueing systems.

Classical Queueing Theory mainly studies the performance of congested dynamic systems when customers follow predetermined behavior. Yet, it lacks the attributes that capture the decision-making process customers undertake when they experience congestion. Often referred to as Rational Queueing (see Hassin and Haviv (2003), Hassin (2016)), my field of interest combines tools of Queueing Theory, together with Game Theory as well as Revenue Management and Optimization, to study the strategic behavior of customers and operators in queueing systems, and, to gain interesting and applicable operational insights.

Customers in service systems act independently in order to maximize their welfare. Yet, each customer's optimal behavior is affected by acts taken by the system operator and/or by other customers. The result is an aggregate "equilibrium" pattern which may not be optimal from the point of view of society as a whole. Of particular interest in the game theoretic approach is the notion of Nash Equilibrium, i.e., a situation (strategy profile) in which no customer has any incentive to change her own decision. Typical questions we attempt to answer in our research are:

1. What are the equilibrium and the socially optimal strategy profiles in the underlying game/decision problem?
2. What managerial steps can be taken such that the socially optimal welfare and/or the optimal revenue for the operator is met in equilibrium?

At the core of our research is studying these matters, but also addressing other issues of importance that arise upon exploration.

This thesis is divided into two complement parts: (a) Observable queueing models (Chapters 2 and 3), in which customers gain information about the system at their arrival, thus, the impact of this information is crucial for their decisions; and (b) Unobservable queueing models (Chapters 4 and 5), where customers choose actions based on how they expect the system to perform given that other customers are also act strategically.

1.1 Observable Queues

Observable queues are queueing systems in which *customers* arrive at a service station, observe its state, and based on this information and other common knowledge, they choose an action that maximizes their welfare. When the action is either to join the system or to balk, and balking does not bring about any gain or loss, rational customers will join as long as their expected total value from joining is nonnegative. This results in *threshold* joining – customers join only in positions smaller than some predetermined threshold.

The next two chapters 2 and 3 are based on Hassin and Snitkovsky (2019) and deal with systems of observable queues. Specifically, in Chapter 2, we introduce and provide general sufficient conditions for a common economic phenomena in observable queueing-models we refer to as *Naor's Inequality*, namely, the monopolist's tendency to overpricing in service systems. In Chapter 3 we apply our results derived in Chapter 2 to many concrete models. All in all, we use these to

1. settle a conjecture and/or prove Naor's inequality in 10 different models (discussed in Adler and Naor (1969), Boudali and Economou (2012), Hasenbein and Chen (2016), Li and Han (2011), Sun and Li (2012), Sun et al. (2018), Wang et al. (2014), Zhang et al. (2014), §3.5,§3.7) and in some special cases of 3 other models (discussed in D'Auria and Kanta (2015), Kerner (2011) and Kim and Kim (2016));

2. generalize theoretical results for Naor's inequality obtained in 4 previous papers (Economou and Kanta (2008a, 2011), Knudsen (1972), Naor (1969) and Simonovits (1976)).

Finally, we provide two examples of models that do not satisfy the inequality.

1.2 Unobservable Queues

In the previous chapters we assume that customers select their actions based on the observed state of the system, hence their strategy is a mapping between states and actions. In the chapters hereafter, we study models where customers cannot observe the queue prior to their actions. The basic unobservable M/M/1 queue were first studied by Edelson and Hildebrand (1975) In terms of analysis, the fundamental difference between the observable queue model, as introduced by Naor (1969), and the unobservable one by Edelson and Hildebrand (1975), is that the latter requires a game theoretic analysis, while in the former, a customer's decision is independent of other customers' actions (yet depends on the observed state).

Motivated by applications of CRNs and Last-Mile Delivery services, Chapter 4 studies a queueing network composed of separated service facilities – one unbounded-capacity queue and one loss system. Time-sensitive customers have to decide weather to join the queue (whose length is not known upon arrival, hence the model is unobservable) or to try to join the loss system, at the risk of being rejected. A corresponding cost-reward structure yields a trade-off between the two options, and the equilibrium strategy profile is analyzed. Comparing the equilibrium and the socially optimal strategies, we arrive at the paper's key result: Contrary to intuition generated by former theoretic results, customers may choose the pricey option more than what is socially preferred. This chapter is based on a paper by Hassin and Snitkovsky (2017).

In Chapter 5 we study a model of an unobservable, single-server priority queue where customers bid for priority. Customers bidding strategy reflect their need for faster service, but is also affected by a behavioral component – i.e., a social norm. This model is motivated by patrons tipping-behavior at service systems, and is based on a paper by

Debo and Snitkovsky (2018). Our model provides answers to the following, fundamental questions: (1) Is adding a social norm sufficient to induce all customers to tip? (2) How does a social norm impact tipping wage and demand? In the context of the raising popularity of restaurants adopting a no-tipping policy, we compare a business model with a service fee versus one with a tip.

PART I: Observable Queues

Chapter 2

Social and Monopoly Optimization in Observable Queues

2.1 Background and Motivation

Economists consistently argue that monopolies are undesirable, because a monopolist restricts production from what a competitive industry would do, under-exploiting the resource in the market, thereby violating economic efficiency. By contrast, if a shared resource is offered free of charge to self-serving individualistic consumers, the collective action often leads to depletion due to overuse. In queueing systems, the underlying product is the service, whose quality usually decreases with system congestion, which in turn, determines the effective demand. This gives rise to a welfare-optimization problem of admission control, which is implemented by imposing an entrance fee. An advantage in studying pricing in strategic queueing systems, is that customers' strategic considerations, alongside the utility structure and the fee, utterly dictate the demand, without having to assume an exogenous demand function. In this work, we suggest a unified approach for studying the aforementioned economic phenomena in the context of *Observable Queues*.

Naor (1969) studied an observable M/M/1 queue with risk-neutral customers who choose between joining or balking. Naor defines three different threshold strategies: The first is the individually optimal (or *equilibrium*[1]) threshold, n_e, which is the threshold followed by customers who join if and only if their expected value from joining is nonnegative.

[1] in dominating strategies

The second is the *socially optimal* threshold, n_o, which is the threshold that maximizes aggregate social welfare per unit time.

The third threshold strategy is derived as follows; Consider a toll-collecting profit-maximizing agency (a *monopolist*) – this agency seeks to impose a fixed toll which maximizes the rate of payments to the server. When a toll is imposed, strategic customers join as long as their surplus (net of toll) is nonnegative, resulting in a joining threshold corresponding with the toll chosen. Moreover, the monopolist chooses the toll such that for the induced threshold n_m, the customer joining in position n_m is indifferent between joining and balking (otherwise the price can be increased without affecting demand). Hence, the problem can be viewed as searching for an optimal threshold, n_m.

The key finding in Naor (1969) is that $n_m \leq n_o \leq n_e$, namely, the effective demand in monopoly is less than the socially optimal demand, which is less than that in equilibrium. Throughout this work we will refer to this three-part relation as *Naor's inequality*.

The part $n_o \leq n_e$ in Naor's inequality is fairly intuitive, and results from a fundamental externalities-based argument (see Proposition 2.3): The joining of a customer may increase the joining position of future customers, which in turn translates into cost. Joining position n_e (or higher) is non-beneficial for the individual, let alone for whole society. This economic phenomenon can be viewed as what was named by Hardin (1968) as the *tragedy of the commons* – under the conditions of scarcity, selfish consumers naturally impose costs on the society. As consumers ignore these negative externalities, they will tend to over-exploit the resource. We briefly discuss this part of the inequality and elucidate this well-known result.

The relation $n_m \leq n_o$, however, need not hold in general, and when it does, it is not easily justified by a simple ramification of an externalities-like argument. Yet, we show, that all the models satisfying $n_m \leq n_o$ in the literature share the following fundamental property: As the monopolist increases the threshold, customers' share of the total welfare increases. For thresholds greater than n_o, the total welfare decreases, hence the monopolist's share decreases too (see Proposition 2.5). Still, proving that customers' share increases with the threshold can be complex: It relies on the simultaneous impacts that changing the threshold makes on the price, the joining probability, and the system's congestion.

Following Naor's work, we discuss sufficient conditions that establish the inequality $n_m \leq n_o$ in a general queueing setup. Later, in Chapter 3 we express these conditions' substance by applying them to a vast range of examples.

2.2 Literature Review

This work builds on a line of research started with a paper by Naor (1969), considering strategic customers in an M/M/1 system. The fundamental relation $n_o \leq n_e$ in Naor's model and its extensions is studied extensively in the literature. Restricted to threshold strategies, many of Naor's extensions and generalizations deal with non-exponential service distribution[2]. Johansen and Stidham Jr. (1980) address the inequality $n_o \leq n_e$ in GI/G/1 queues with a generalized cost structure, in finite and infinite horizon[3]. Adler and Naor (1969) show a continuous analog to $n_o \leq n_e$ in terms of workload, in M/D/1 queues with linear waiting cost. In general, for non-exponential service distributions, an equilibrium in pure threshold strategies need not exist (see, for instance Kerner (2011), Altman and Hassin (2002)). We discuss this observation thoroughly in §3.10, studying Naor's inequality in M/G/1 systems.

Assuming exponential service, the inequality $n_o \leq n_e$ has been studied by Lippman and Stidham Jr. (1977) and Stidham Jr. (1978) in a general cost structure, with finite and infinite horizon. However, it is not always true that the welfare optimizing control is of threshold type. Mendelson and Yechiali (1981) show, in a GI/M/1 system with linear waiting costs, that *conditional acceptance* strategies, which allow the reneging of the last customer in the queue, may increase social welfare. When reneging is prohibited, as in the present paper, Yechiali (1971, 1972) shows how to compute n_o in GI/M/s systems. Xu and Shanthikumar (1993) and Wang Wang (2016) deal with calculating n_o using the so called *dual approach* method[4].

Some papers, such as Hassin (1985) and Haviv and Oz (2016), focus on designing mechanisms for system regulation, such that the socially optimal threshold is met in equilibrium, that is, $n_o = n_e$. In a related work, Kim et al. (2011) analyze the *last-come first-priority* regime in M/M/s heterogeneous-servers systems and show it implies $n_o = n_e$. We discuss examples of G/M/s queues with homogeneous and heterogeneous servers in §3.2. Other papers we refer to later concerning the relation $n_o \leq n_e$ include D'Auria and Kanta (2015) and Kim and Kim (2016) for tandem queues, and Li and Han (2011) for queues with breakdowns.

Being less intuitive, the complement part of Naor's inequality, $n_m \leq n_o$, has received less attention in the theory. Knudsen (1972) generalizes Naor (1969), showing that $n_m \leq n_o$ for M/M/s queues where customer utility is non-increasing and *concave* in the joining position. Simonovits (1976) also generalizes Naor (1969), but does not generalize Knudsen (1972), showing a similar result for GI/M/s queues and *linear* non-increasing

[2]For a survey on strategic queueing with non-exponential service distribution see Hassin (2016)§2.1.2.
[3]For a survey on optimal admission control in queues see Stidham Jr. (1985).
[4]For more information, see Hassin (2016)§2.2.

utility. Economou and Kanta (2011) prove that $n_m \leq n_o$ for a FCFS Orbit queue with linear non-increasing utility, and an analogous result in Economou and Kanta (2008a) for Naor's model with compartmented waiting space. Other papers (e.g., Boudali and Economou (2012), Hasenbein and Chen (2016), Sun and Li (2012), Sun et al. (2018), Wang et al. (2014), Zhang et al. (2014)) provide numerical support for $n_m \leq n_o$ in several different models.

The inequality $n_m \leq n_o$ can be interpreted as the monopolist's tendency to reduce effective demand by overpricing its service. De Vany (1976) considers an observable queue with an exogenous, price-dependent arrival rate, and shows the monopolist always overprices the service. However, the arrival rate being dependent on the price suggests that customers are heterogeneous, which is inconsistent with the analysis[5]. Thus, the explanation by De Vany (1976) is not valid when the effective demand is endogenously formed, as in this work, due to customers decisions.

2.3 Model Description

First, we introduce the queueing process: Consider a service system with a single queue in which the waiting slots are totally ordered. There is an exogenous arrival process of *potential* homogeneous customers to the system which is a general stationary counting process[6]. The queue is observable: Upon arrival to the system, a customer is offered a joining position, which is the total number of customers in the system if that customer joins[7]. Given the offered position, the customer then chooses either to join that position or to balk.

Definition 2.1. A pure threshold joining strategy $n \in \mathbb{N}$[8] is a joining rule of the form: "Join if and only if the offered position is not (strictly) greater than n."

It follows that when customers adopt an n-threshold strategy, the positions likely to be offered (i.e., with positive probability) take values on the set $\{1, \ldots, n+1\}$, and when position $n+1$ is offered the customer balks.

We further assume that the sequence of offered positions, embedded at arrival instants, converges in distribution for every threshold strategy n, and we denote its limit by $Q^{(n)}$. In other words, the random variable $Q^{(n)}$ denotes the offered position for an arriving customer in steady-state, assuming the entire population follow the threshold

[5] For criticism, see Chen and Frank (2004).

[6] We allow for group arrivals (that is, the possibility of zero inter-arrival times) as long as the order of customers within a group is defined.

[7] A slight exception is made in §3.2.2, §3.8 and §3.9.

[8] Taken as the set of *positive* integers, $1, 2, 3, \ldots$

n. Naturally, we assume that for all n, $Q^{(n)} \leq_{\text{st}} Q^{(n+1)}$ (where '$\leq_{\text{st}}$' denotes usual stochastic-order domination), namely, increasing the threshold causes higher congestion in steady state.

Next, we define the utility structure. Every joining position is associated with a value which is *independent of customers joining strategy*: For $k \in \mathbb{N}$, let $u(k)$ be the expected utility for a customer who joins the system in position k. A balking customer, w.l.o.g, receives zero utility. Thus, utility-maximizing customers join position k if and only if $u(k) \geq 0$, otherwise they balk. We emphasize that the assumption that $u(k)$ is a function of sheer position implies that $u(k)$ can be assessed without having to conjecture the other customers' strategy, in which case we say that $u(k)$ is *strategy independent*.

To simplify the notation, we let $B_n = \{Q^{(n)} = n+1\}$, and denote its complement event by $J_n = \{Q^{(n)} < n+1\}$. Recall that $Q^{(n)}$ represents the stationary offered position embedded in arrival instants for threshold n. Thus, the event B_n corresponds to the $b$alking of an arbitrary arriving customer, whereas J_n corresponds to the case she $j$oins. When customers follow threshold n, an arriving customer's utility, $S^{(n)}$, is a random variable that depends on $Q^{(n)}$ and can be expressed as

$$S^{(n)} = u\big(Q^{(n)}\big) \cdot \mathbf{1}_{J_n}, \tag{2.1}$$

where the random variable $\mathbf{1}_{J_n}$ is the indicator function of the event J_n.

When the service provider is a non-discriminating monopolist, her revenue per customer is the toll levied when that customer joins, and zero otherwise. The monopolist collects the same toll from every joining customer, and customers join as long as the toll is not greater than their expected net benefit. As explained in Hassin and Haviv (2003)§2.4, the optimal toll levied by the monopolist is either of the form $u(k)$ for some $k \in \mathbb{N}$, or $\lim_{k\to\infty} u(k)$ (otherwise, by increasing the toll, the monopolist can increase profit without affecting the admission process). In other words, to induce a threshold n, the monopolist charges an admission fee equal to the utility of the customer who enters in position n. We define the random variable $M^{(n)}$ as the monopolist's revenue per arriving customer:

$$M^{(n)} = u(n) \cdot \mathbf{1}_{J_n}. \tag{2.2}$$

The expected social welfare and the monopolist's expected profit, per customer, as functions of the threshold n, are $\mathrm{E}\big(S^{(n)}\big)$ and $\mathrm{E}\big(M^{(n)}\big)$, respectively. We denote

$$n_e = \max\{n \in \mathbb{N} \mid u(n) \geq 0\}, \quad n_o = \arg\max_{n\in\mathbb{N}} \mathrm{E}\big(S^{(n)}\big), \quad n_m = \arg\max_{n\in\mathbb{N}} \mathrm{E}\big(M^{(n)}\big), \tag{2.3}$$

when the maximum is attained, otherwise ∞. When finite, we assume n_o (similarly, n_m) is unique, otherwise we simply take the smallest n_o (similarly, n_m) such that the function $\mathrm{E}\big(S^{(n)}\big)$ (similarly, $\mathrm{E}\big(M^{(n)}\big)$) is maximized.

We further denote the random variable $D^{(n)} = S^{(n)} - M^{(n)}$. An interpretation for $D^{(n)}$ is that it represents a random customer's net surplus (utility minus fee) in the monopolistic system of threshold n, that is, with admission fee $u(n)$.

Definition 2.2. The system satisfies *Naor's inequality* if $n_m \leq n_o \leq n_e$.

In what follows, we focus on deriving sufficient conditions for Naor's inequality.

2.4 Basic Results

The following Propositions 2.3 and 2.4 are well known in the literature in various versions and phrasings (see e.g., Hassin and Haviv (2003)§2.1), therefore their proofs are omitted. The results achieved hereafter build on these propositions, therefore we find it necessary to reiterate them while using the setting and definitions presented above:

Proposition 2.3. *If the sequence* $\{u(k)\}_{k=1}^{\infty}$ *is non-increasing, then* $n_o \leq n_e$.

This result is due to the negative externalities generated when customers join: Suppose that an arriving customer joins position n_e+1 or higher. Clearly, she will incur negative utility. Since $\{u(k)\}_{k=1}^{\infty}$ is non-increasing and congestion increases with the joining of new customers ($Q^{(n)} \leq_{\mathrm{st}} Q^{(n+1)}$), her joining may only reduce the utility of arriving future customers. Customers who arrived before her are not affected. Thus, overall, this customer's joining causes a strict decrease in social welfare, meaning $n_o \leq n_e$.

Proposition 2.4. *If the sequence* $\{u(k)\}_{k=1}^{\infty}$ *is non-increasing, then* $n_m \leq n_e$.

This statement is rather obvious; By the definition of n_e, $u(n_e + 1) < 0$. To induce a threshold $n > n_e$, the monopolist should charge price $u(n) < 0$, implying negative expected revenue.

2.4.1 Fundamental Result

The next statement, although intuitive, is new and not mentioned in the literature. This proposition is, in fact, the core building block in this chapter, in the sense that all of our results concerning Naor's inequality rely on it:

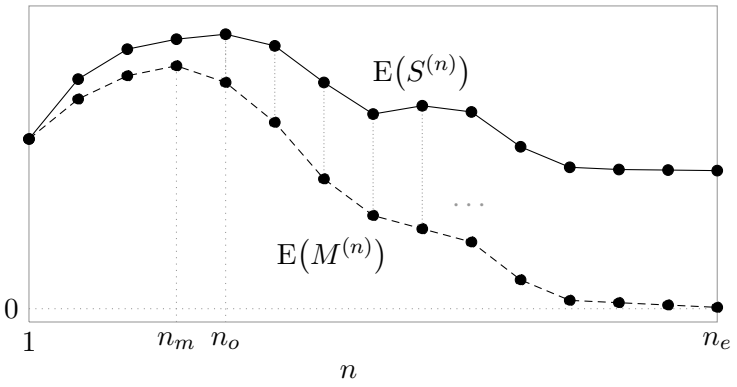


FIGURE 2.1: A graphical interpretation of Proposition 2.5. The vertical dotted lines reflect the sequence $\mathrm{E}\big(D^{(n_o)}\big), \mathrm{E}\big(D^{(n_o+1)}\big), \ldots$, whose lengths are increasing, which implies $n_m \leq n_o$.

Proposition 2.5. *If the sequence* $\{u(k)\}_{k=1}^{\infty}$ *is non-increasing, and* $\mathrm{E}\big(D^{(n)}\big) \leq \mathrm{E}\big(D^{(n+1)}\big)$ *for all* $n \in [n_o, n_e]$*, then* $n_m \leq n_o$.

The proof of Proposition 2.5 is in Appendix A.1. A graphical interpretation is depicted in Fig. 2.1.

The meaning of the condition $\mathrm{E}\big(D^{(n)}\big) \leq \mathrm{E}\big(D^{(n+1)}\big)$ is that when the threshold increases above n_o, customer expected surplus in the monopolistic system increases too. However, n_o maximizes the overall social welfare and increasing the threshold above n_o reduces overall social welfare. Therefore it must be that the monopolist's share decreases. Thus, the monopolist chooses a threshold not greater than n_o.

In contrast to Propositions 2.3 and 2.4 which only depend on the structure of the utility, the condition of Proposition 2.5 depends both on the utility structure and on the queueing dynamics. This can be easily seen via the definition of $\mathrm{E}\big(D^{(n)}\big)$. We shall see later, in §3.11, that unlike in Propositions 2.3 and 2.4, $\{u(k)\}_{k=1}^{\infty}$ being non-increasing is not a sufficient condition for $n_m \leq n_o$.

Verifying the condition of Proposition 2.5 can sometimes be challenging. The reason being that increasing the threshold in the monopolistic system (which is done by reducing the price) may have contradicting effects on customer surplus; On the one hand, it reduces both the balking probability and the admission fee (which in turn increase customer expected surplus). On the other hand it involves joining a longer, more congested queue, which decreases customer surplus. Despite the complexity, we shall show how several common features, shared among many models, can be utilized in verifying the condition.

2.5 Key Assumptions

Our goal is to prepare the ground for introducing Proposition 2.10, which provides a more easily-verified equivalent to Proposition 2.5. We precede the analysis by introducing, and later on discussing, two basic assumptions:

(A-i) The expected utility for a customer is a strategy-independent non-increasing function of her joining position, i.e., $u(1) \geq u(2) \geq u(3) \geq \dots$

(A-ii) For every threshold[9] $n \in \mathbb{N}$, the stochastic ordering $Q^{(n)} \leq_{\text{st}} Q^{(n+1)} \leq_{\text{st}} Q^{(n)} + 1$ holds.

An equivalent representation of Assumption (A-ii) is the following (see Shaked and Shanthikumar (2007)§1.A.1): For every $n \in \mathbb{N}$ and $k \in \{1, \dots, n+1\}$,

$$\Pr\big(Q^{(n)} > k\big) \leq \Pr\big(Q^{(n+1)} > k\big) \leq \Pr\big(Q^{(n)} > k-1\big).$$

Unless stated differently, we assume henceforth in the analysis that Assumptions (A-i) and (A-ii) hold. To avoid trivialities, we further assume $u(1) > 0$, otherwise customers have no incentive to join, implying $n_m = n_o = n_e = 0$.

2.5.1 Discussion of Assumption (A-i)

Assumption (A-i) is the most fundamental property in the discussion of threshold strategies, and is in accordance with the assumptions made by Knudsen (1972) and Simonovits (1976). This assumption implies that customers, as they join (thereby increasing the queue length), induce negative externalities on future arriving customers. These externalities are the key explanation for the segment $n_o \leq n_e$ in Naor's inequality (see Proposition 2.3). However, n_e might be strictly larger than n_o if there exists k such that $u(k) < u(k+1)$.[10]

The most prevalent example for a strategy-independent utility is given by Naor (1969), where customer waiting cost is linear, and $u(k) = R - \frac{Ck}{\mu}$, for positive constants R, C and μ. The optimal joining criterion for a customer is to join position k if and only if $R - \frac{Ck}{\mu} \geq 0$, regardless of the decisions of others, meaning this is a *dominant* strategy. Consider, on the contrary, the model with linear waiting cost and Egalitarian Processor-Sharing service regime, studied by Altman and Shimkin (1998). An arriving

[9] Once n_o and n_e are defined (see (2.3)), this assumption can be relaxed to "for every $n \in [n_o, n_e]$".
[10] See, for example Hassin and Haviv (2003) §1.5.2.

customer's sojourn time is a function of both the observed number of customers and of future customers' joining strategy. Hence, her utility is a function of those two as well. Models such as Altman and Shimkin (1998) do not comply with Assumption (A-i).

Intuitively, it may be argued that the assumption that $u(k)$ is strategy independent is a direct consequence of the FCFS regime. Particularly, in a G/M/1 FCFS system with customer utility depending only on waiting time, $u(k)$ is indeed strategy independent. Yet, assuming FCFS is not required in our general setup, because once we have defined $u(k)$ for every $k \in \mathbb{N}$, the relation between the utility and the actual waiting time is no longer relevant. For instance, if we assume in Naor's model (Naor (1969)) that a customer, regardless of her actual waiting time, receives $R - \frac{Ck}{\mu}$ *immediately* after joining position k, then the analysis remains as in the original model, under any work-conserving[11] service regime.

It should be mentioned in this context that there exist models studying FCFS queues where the utility is not strategy independent (see, for example, Burnetas and Dimitrakopoulos (2018)). Conversely, some models, like the one studied by Economou and Kanta (2011) (see §3.9), and the one in §3.7, are such that although the service regime is not FCFS, $u(k)$ is still strategy independent.

2.5.2 Discussion of Assumption (A-ii)

Assumption (A-ii) relates to the queueing process, but not to the utility structure. We use this assumption in the analysis as a tool for comparing the same system operating under different thresholds, n and $n+1$. To put into words, Assumption (A-ii) means that as one increases the threshold in the system from n to $n+1$, the (random) number of customers in the system grows (stochastically), but does not grow as much as having one extra customer in the system all the time. Despite this property being intuitive and applicable for a broad range of common queueing models[12], to the best of our knowledge, there is no previous reference to it in the queueing literature.

Roughly speaking, by introducing Assumption (A-ii), we are able to bound the amount by which congestion grows in the system when the threshold increases by one. It allows us to assume w.l.o.g that when a customer is offered position k in a system with threshold n, then a customer arriving at the same time to the system with threshold $n+1$ is offered either position k or $k+1$. Relaxing this assumption opens the door to pathological examples, where the systems with thresholds n and $n+1$ are substantially different, so

[11] Assuming a non-work-conserving regime in Naor's model can change the stationary distribution of $Q^{(n)}$ therefore work conservation is required.

[12] including G/M/s, M/G/1, models incorporating vacations, catastrophes, retrials and more (see §3).

much that they resemble two separate, unrelated and incomparable processes. Lacking a solid basis for the comparison of the system operating under different thresholds, the attempt of searching for a meaningful conditions for Naor's inequality seems futile. An example for a queueing system where Assumption (A-ii) does not hold is given in §3.11.2, clarifying why this assumption is crucial in proving Naor's inequality.

When the distribution of $Q^{(n)}$ is given explicitly, Assumption (A-ii) can be verified algebraically. Of course, most challenging are cases where the stationary distribution is not easily expressed, and one has to utilize properties of the queueing process to establish Assumption (A-ii). In particular, in Section §3.10 we show that if the system is M/G/1 then the assumption holds. One very common property that implies Assumption (A-ii) is memoryless (either discrete or continuous) service. Formally:

Proposition 2.6. *In a non-anticipating regime where at any time instant, the residual-service requirements of all customers are i.i.d (memoryless), Assumption (A-ii) holds.*

The proof of Proposition 2.6 is in Appendix A.2.

2.6 Prior Results

Next we mention two results (Propositions 2.7 and 2.8) that are the most general conditions for $n_m \leq n_o$ so far appearing already in the literature. The proofs of these propositions are omitted as they appear in Knudsen (1972) and Simonovits (1976). Regarding the arrival process, we write G to indicate a general stationary processes (inter-arrival times need not be independent), and GI to indicate a renewal processes. When referring to customers' service duration, unless stated otherwise, we assume they are independent.

Proposition 2.7 (Knudsen (1972)§6, Theorem 2)**.** *Assuming (A-i), if the system is an M/M/s, and* $\{u(k)\}_{k=1}^{\infty}$ *is a* concave *sequence, then* $n_m \leq n_o \leq n_e$.

Proposition 2.8 (Simonovits (1976)§5, Proposition 2)**.** *Assuming (A-i), if the system is a GI/M/s, and* $\{u(k)\}_{k=1}^{\infty}$ *is a* linear *sequence, then* $n_m \leq n_o \leq n_e$.

Propositions 2.7 and 2.8 are neither more nor less general than each other. We note that assuming GI/M/s (and in particular, M/M/s), by Proposition 2.6, implies assumption (A-ii), and therefore it is more restrictive than (A-ii). We later introduce Proposition 2.10, which implies, among other things, that Assumptions (A-i) and (A-ii), together with $\{u(k)\}_{k=1}^{\infty}$ being concave, are sufficient conditions for Naor's inequality (see Corollary 2.12), thus, immediately generalizing both Propositions 2.7 and 2.8.

2.7 Analysis

Proposition 2.5 suggests that to study Naor's inequality one has to consider the change in customer surplus in the monopolistic system as the threshold grows. A larger threshold imposes higher congestion in the system, and therefore lower expected utility. Hence, it is of interest to study the probability that the queue observed in an $(n+1)$-threshold system is strictly longer than in the n-threshold system. We define this event rigorously as the event A_n below.

Theorem 1.A.1 in Shaked and Shanthikumar (2007) states that two random variables X and Y satisfy $X \leq_{\text{st}} Y$ if and only if there exists a probability space in which $\Pr(X \leq Y) = 1$. Having posed Assumption (A-ii), we therefore assume, w.l.o.g, that $Q^{(n)}$ and $Q^{(n+1)}$ are defined on the same probability space (i.e., the *coupling* space), such that $Q^{(n)} \leq Q^{(n+1)} \leq Q^{(n)} + 1$ with probability 1, for any specified threshold n. In the coupling space (denoted by $(\Omega, \mathcal{F}, \Pr)$), we define the event

$$A_n = \left\{\omega \in \Omega \mid Q^{(n+1)}(\omega) = Q^{(n)}(\omega) + 1\right\},$$

to immediately obtain

$$Q^{(n+1)} = Q^{(n)} + \mathbf{1}_{A_n}, \tag{2.4}$$

where $\mathbf{1}_{A_n}$ denotes the indicator function of A_n.

Following the construction described in A.1, when the service is exponential, we can provide an intuitive interpretation for the event A_n, based on the notion of the *standby customer* (see Haviv (2013)§4.7.3); In an n-threshold system, suppose that when a customer is offered to join position $n+1$, instead of balking she joins as a standby customer. That is, she obtains service only when there are no other customers waiting for her in the system. By construction, at any time there can be no more than one standby customer. The stationary number of customers (at arrival instants) including the standby customer is given by $Q^{(n+1)}$, and excluding the standby customer it is $Q^{(n)}$. Then, A_n is the event that a moment after an arbitrary arrival, a standby customer is present in the system. In many of our examples (e.g., §3.2, §3.5 and §3.6) we interpret the event A_n this way. Figure 2.2 depicts the dynamics of state transitions in a system with the standby customer. We stress that we use the notion of the standby customer in a context completely unrelated to the utility function. It merely serves as a conceptual tool to validate Assumption (A-ii).

Let $u'(k) = u(k+1) - u(k)$. Since $u(k)$ is non-increasing, $u'(k)$ is nonpositive for all k. The following lemma provides an expression for $\mathrm{E}\big(D^{(n+1)} - D^{(n)}\big)$ in terms of A_n, $Q^{(n)}$ and the function $u'(k)$:

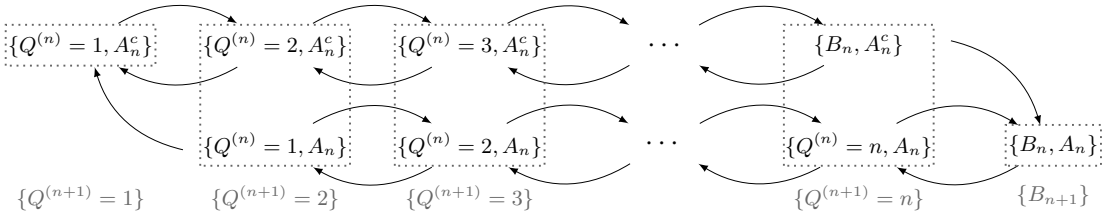


FIGURE 2.2: An examples of a possible queueing process and its transition between offered positions in an n-threhold system with the event A_n. Each event $\{Q^{(n+1)}=k\}_{k=1}^{n}$ and B_{n+1} can be represented as a union of the (disjoint) events listed above it.

Lemma 2.9. *Assuming (A-ii), for every* $n \in \mathbb{N}$,

$$\mathrm{E}\big(D^{(n+1)} - D^{(n)}\big) = \mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n\right) \cdot \Pr(J_n \cap A_n) - u'(n) \cdot \Pr(J_n). \tag{2.5}$$

The proof of Lemma 2.9 is in Appendix A.3.

Equation (2.5) can be explained as follows: Suppose we have two coupled systems, one with threshold $n+1$ (and admission fee $u(n+1)$) and the other one with threshold n (and admission fee $u(n)$). The queue length in the $(n+1)$-system, at any time, is bigger than that in the n-system by no more than 1 (note that in this case, $J_n \subseteq J_{n+1}$). Then, we can express the difference in expected surplus (net of fee) between a customer arriving at the $(n+1)$-system and a customer arriving the same moment at the n-system: When both customers join (the event J_n), then the difference in surplus consists of (a) the difference in fees, $-u'(n)$, and (b) the difference in (gross) utility, which depends on the joining positions:

- If the position in the $(n+1)$-system is bigger by 1 than in the n-system (the event A_n), then the utility difference is $u'\big(Q^{(n)}\big)$.
- If the positions are identical in both systems (the event A_n^c), the utility difference is 0.

In all other cases, there is no difference in the two customers' surplus: When both customers are blocked (the event $\{B_n \cap A_n\}$) they would receive zero net surplus in both systems. If the customer in the $(n+1)$-system joins and the other customer balks (the event $\{B_n \cap A_n^c\}$), then the former must have joined position $n+1$, thus, her net surplus is 0, as well as the latter balking customer. Other cases are impossible as those imply that the difference between the offered positions is 2 or more. Overall, the expected difference in surplus sums up to the right-hand side of Equation (2.5).

The next Proposition 2.10 and its derivatives, Corollaries 2.12 and 2.11 are the main results of this section, as they are used the most extensively in Chapter 3.

Proposition 2.10. *Assuming (A-i) and (A-ii), if*

$$\Pr(A_n \mid J_n) \leq \frac{u'(n)}{\mathrm{E}\left(u'\left(Q^{(n)}\right) \mid J_n \cap A_n\right)}, \quad \forall n \in [n_o, n_e], \tag{2.6}$$

then $n_m \leq n_o \leq n_e$.

The proof of Proposition 2.10 is in Appendix A.4.

We can further simplify the conditions in Proposition 2.10 when more properties of $u(k)$ are known in addition to monotonicity (Assumption (A-i)):

Corollary 2.11. *Assuming (A-i) and (A-ii), if* $\{u(k)\}_{k=1}^{\infty}$ *is a convex sequence, and* $\Pr(A_n \mid J_n) \leq \frac{u'(n)}{u'(1)}$ *for all* $n \in [n_o, n_e]$, *then* $n_m \leq n_o \leq n_e$.

The proof of Corollary 2.11 is in Appendix A.5.

The term $\Pr(A_n \mid J_n)$ represents the probability that a joining customer in the $(n+1)$-system encounters higher congestion than a customer joining the n-system at the same time. Thus, we can consider this term as a measure of the marginal growth in congestion when the threshold increases. Similarly, the right-hand side of Equation (2.6) and the term $\frac{u'(n)}{u'(1)}$ relate to the marginal decay in utility. Proposition 2.10 and Corollary 2.11 state that if the rate at which congestion grows in the system (with respect to the threshold) is sufficiently slow compared to the rate at which utility decays, then Naor's inequality is satisfied. This is intuitive, because unless congestion varies significantly between the n- and the $(n+1)$-system, then, due to the reduction in fee and in the balking probability, a customer would be better off (in terms of expected surplus) joining the latter system, and, following Proposition 2.5, Naor's inequality holds.

In the general case, when verifying the conditions of Proposition 2.10, one has to consider changes in the queueing process (i.e., the distribution of $Q^{(n)}$) with respect to the threshold. Interestingly, when $u(k)$ is concave, then, by Corollary 2.12 below, Naor's inequality holds regardless of the distribution of $Q^{(n)}$.

Corollary 2.12. *Assuming (A-i) and (A-ii), if* $\{u(k)\}_{k=1}^{\infty}$ *is a concave sequence, then* $n_m \leq n_o \leq n_e$.

The proof of Corollary 2.12 is in Appendix A.6.

Corollary 2.12 elucidates why the result established in Proposition 2.10 is more general than both Proposition 2.7 (by Knudsen (1972)) and Proposition 2.8 (by Simonovits

(1976)). The intuition behind Corollary 2.12 is the following: Suppose that a customer in the monopolistic n-system is offered to join in position $k \leq n$ (and pay a fee $u(n)$). By Assumption (A-ii) a customer joining the $(n+1)$-system at the same time would be offered either position k or $k+1$ (and pay a fee $u(n+1) \leq u(n)$). But since $u(k)$ is concave decreasing, $u(k) - u(n+1) \geq u(k+1) - u(n+1) \geq u(k) - u(n)$ (for every $k \leq n+1$), meaning that in either case, the surplus in the $(n+1)$-system is greater than in the n-system. Thus, by Proposition 2.5, Naor's inequality holds.

2.7.1 Secondary Results

As discussed, the conditions in Proposition 2.10 and Corollary 2.11 (when concavity is not assumed) depend both the distribution of $Q^{(n)}$ and the structure of $u'(n)$ (for every $n \in [n_o, n_e]$). We next present the following expansions, that simplify the analysis required in order to use Proposition 2.10 and Corollary 2.11. These results are used in §3.5 and §3.6.

Lemma 2.13. *Assuming (A-i) and (A-ii), for every $n \in \mathbb{N}$,*

(i) $\Pr(A_n) = \mathrm{E}\big(Q^{(n+1)}\big) - \mathrm{E}\big(Q^{(n)}\big)$,

(ii) $\Pr(A_n \mid J_n) = \frac{\mathrm{E}\big(Q^{(n+1)}\big) - \mathrm{E}\big(Q^{(n)}\big) - \Pr(B_{n+1})}{\Pr(J_n)}$,

(iii) for $k = 1, 2, \ldots, n+1$,

$$\Pr\big(Q^{(n)} = k, A_n\big) = \sum_{i=1}^{k} \Big(\Pr\big(Q^{(n)} = i\big) - \Pr\big(Q^{(n+1)} = i\big)\Big). \tag{2.7}$$

The proof of Lemma 2.13 is in Appendix A.7.

The next result (Lemma 2.14) exploits two common features of queueing systems: (a) The sequence of the state probabilities $\{\Pr\big(Q^{(n)} = k\big)\}_{k=1}^{n+1}$ is decreasing for every threshold n (this is expected when the service rate is larger than the arrival rate); (b) When increasing the threshold by one, the stationary probability of each state (except for the newly added state) decreases, i.e, $\Pr\big(Q^{(n+1)} = k\big) \leq \Pr\big(Q^{(n)} = k\big)$ for all $k = 1, \ldots, n+1$. This result is used in the proof of Lemma 3.6.

Lemma 2.14. *Assuming (A-i) and (A-ii) and given a threshold $n \in \mathbb{N}$, if for all $k \leq n+1$,*

(i) $\Pr(B_n) \leq \Pr\big(Q^{(n)} = k\big)$, *and*

(ii) $\Pr\big(Q^{(n+1)} = k\big) \leq \Pr\big(Q^{(n)} = k\big)$,

then $\Pr(A_n \mid J_n) \leq \Pr(A_n)$.

The proof of Lemma 2.14 is in Appendix A.8.

2.8 Concluding Remarks

This work deals with Naor's inequality, in observable queues where joining-customers utility is decreasing with the joining position. The validity of the common relation $n_m \leq n_o$ depends on the relation between the utility function and the queueing process, and may not hold in general. We establish sufficient conditions for $n_m \leq n_o$, based on coupling the system when customers' threshold is n with the same system when the threshold is $n+1$. When service demand is exponential, this coupling can be constructed by applying a service regime with a standby customer. We conjecture that a similar construction can also be used in proving that Assumption (A-ii) holds not only when the service demand is memoryless, but also when it is *new-better-than-used* (NBU). We further show a different proof technique that obtains Assumption (A-ii) in an M/G/1. A natural extension is generalizing this technique for an M/G/s system.

Chapter 3

Model Examples for Naor's Inequality

In this chapter we survey several model examples, most of which where previously discussed in the literature in the context of Naor's inequality. Through these models we demonstrate how to apply our results derived in Chapter 2, namely Proposition 2.10 and its derivatives, Corollaries 2.11 and 2.12, in order to prove Naor's inequality. Additionally, we show by two different examples how relaxing our assumptions in Chapter 2 may result in violating Naor's inequality.

3.1 Literature Review

Following Knudsen (1972), Naor (1969) and Simonovits (1976), we introduced in Chapter 2 conditions (Proposition 2.10) that imply the full Naor's inequality, $n_m \leq n_o \leq n_e$. Our result is more general than Economou and Kanta (2011), Knudsen (1972), Naor (1969) and Simonovits (1976), and its strength is expressed in many aspects: It neither assumes concave customer utility nor memoryless service, and it is not restricted to renewal arrival processes. We demonstrate how the result applies to many concrete models, one of which is the Abandonment Model (see §3.5) inspired by Garman Garman (1976), where traders arrive at a market and place bidding orders that last for some stochastic 'lifetime'. More applications discussed in Chapter 3 include: queues with parameter uncertainty (based on Hasenbein and Chen (2016), see §3.2.2); tandem queueing networks (based on D'Auria and Kanta (2015), Kim and Kim (2016), see §3.3); queues with breakdowns (based on Li and Han (2011), see §3.4); queues with catastrophes (based on Boudali and Economou (2012), see §3.6); compartmented M/M/1 (based on Economou and Kanta (2008a), see

§3.8); orbit queues (based on Zhang et al. (2014), Economou and Kanta (2011), see §3.9); and M/G/1 queues (based on Adler and Naor (1969), Kerner (2011), see §3.10);

Contrarily to Naor's result, for heterogeneous customers, Edelson and Hildebrand (1975) show that a monopolist may *over-exploit* the system (i.e., under-price the service). However, a question so far remained unanswered is, for *homogeneous* customers, is the same condition for $n_o \leq n_e$ also sufficient for $n_m \leq n_o$? We disprove this conjecture in §3.11 by two examples in which $n_o < n_m$, namely, the monopolist over-exploits the system.

3.2 G/M/s with Convex Waiting Cost

Consider a G/M/s FCFS queue with service rates $\mu_1, \mu_2, \ldots, \mu_s$. We discuss both cases of homogeneous and heterogeneous servers. Assume the utility of service after waiting t units of time in the system is given by $R - c(t)$, where $R > 0$ is some fixed reward and $c(t)$, the waiting cost function, is convex. Customers observe the queue length upon arrival, and then decide whether to join or balk.

Prior to analyzing the utility for a customer we introduce the following Lemma 3.1 and Corollary 3.2 that are later used in the example. This results relate to the interaction between the cost when expressed as a function of (continuous) waiting time, and customer utility given the (discrete) offered position.

Lemma 3.1. *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d nonnegative random variables, let Y be a nonnegative random variable independent of all $\{X_i\}_{i=1}^{\infty}$ and let $g(x)$ be a convex function. Define $S_0 = Y$ and $S_n = Y + \sum_{i=1}^{n} X_i$, $n = 1, 2, \ldots$, then the sequence $\{\mathrm{E}\left(g(S_n)\right)\}_{n=0}^{\infty}$ is convex.*

The proof of Lemma 3.1 is in Appendix B.1. As a direct result we have:

Corollary 3.2. *Let $\{X_i\}_{i=0}^{\infty}$ be a sequence of i.i.d nonnegative random variables, and let $g(x)$ be a convex function. Define $S_n = \sum_{i=0}^{n} X_i$, then the sequence $\{\mathrm{E}\left(g(S_n)\right)\}_{n=0}^{\infty}$ is convex.*

This is a particular case of Lemma 3.1 setting $Y = X_0$.

3.2.1 Homogeneous Servers

Suppose that $\mu_1 = \mu_2 = \cdots = \mu_s = \mu$, so the system is a standard G/M/s, and by Proposition 2.6, Assumption (A-ii) holds.

Consider a customer who joins the system in position k. If $k \leq s$, her time spent in the system includes only the service time, which is exponentially distributed with rate μ. If $k > s$, her waiting time in the queue (excluding self service) is Erlang distributed with shape parameter $k - s$ and rate $s\mu$, and adds up to the time spent in service. Thus, conditioned on the queue length, the total time spent in the system, W_k, is a sum of $(k-s)^+$ i.i.d nonnegative random variables plus an independent random variable, therefore, by Lemma 3.1, $\mathrm{E}(c(W_k))$ is convex as a function of k. Denote by $u(k)$ the expected utility of that particular customer, then $u(k) = R - \mathrm{E}(c(W_k))$. It follows that $\{u(k)\}_{k=1}^{\infty}$ is a non-increasing and concave sequence. From Corollary 2.12, $n_m \leq n_o \leq n_e$. We therefore formulate the following observation:

Corollary 3.3. *In the observable G/M/s (homogeneous servers) with fixed service reward R and convex increasing time cost function $c(t)$, $n_m \leq n_o \leq n_e$.*

Special cases of this model were considered in the following papers:

- Naor (1969), considers $s = 1$, Poisson arrivals and a linear cost function.
- Knudsen (1972)§7, considers Poisson arrivals and a piecewise linear and convex cost function.
- Simonovits (1976) considers general independent interarrival-times and a linear cost function.
- Sun and Li (2012) consider $s = 1$, Poisson arrivals and $c(t) = C \cdot t^m$ for $m = 1, 2, 3$ and $C > 0$. They provide numerical evidence to Naor's inequality for the specified model. In fact, by Corollary 3.3 Naor's inequality holds for every $m \geq 1$.
- Wang et al. (2014) consider $s = 1$, Poisson arrivals and upon waiting t time units, a customer receives $(V - Ct) - b(V - Ct)^2$, for some nonnegative constants C, V and $b \leq \frac{1}{2}$. This is a special case of Corollary 3.3 with $R = V(1 - bV)$ and $c(t) = (1 - 2b)Ct + bC^2t^2$.

In all these special cases the inter-arrival times are i.i.d. The following example is devoted to an interesting related model studied by Hasenbein and Chen (2016), where inter-arrival times are *not* independent.

3.2.2 Naor's Model with Unknown Arrival Rate

Similar to Naor (1969), Hasenbein and Chen (2016)[1] consider a single-server, exponential service, FCFS queue, but the arrival rate is a random variable Λ with distribution over

[1] Hasenbein and Chen (2016) study both an observable and an unobservable versions of the model.

the support $[0, \overline{\lambda}]$ and cumulative distribution function $F(\lambda)$. Given the realization $\Lambda = \lambda$, the arrival process is Poisson with rate λ. As in Naor's model, they consider linear customer waiting cost $C > 0$ per unit time, which implies $u(k) = R - \frac{Ck}{\mu}$. However, the limiting average utility for a customer here is a function of the random variable Λ. Note that inter-arrival times are not independent (in the sense that by sampling the time between arrivals one can make estimations of the realization of Λ). Hasenbein and Chen observed by numeric examples that Naor's inequality holds.

To adopt our previous notation we reformulate the problem as follows: Given any realization $\Lambda = \lambda$, let $q_\lambda = \frac{\lambda}{\overline{\lambda}}$, and let $X_\lambda^{(n)}$ denote the offered position in an M/M/1/n system in steady state, when the arrival rate is Poisson with rate λ. We assume the overall arrival rate to the system is $\overline{\lambda}$. Each arriving customer is offered independently with probability q_λ to join in position $X_\lambda^{(n)}$, otherwise, with probability $1 - q_\lambda$ she is offered position $n+1$. This implies that the overall balking probability is $q_\lambda \Pr\big(X_\lambda^{(n)} = n+1\big) + 1 - q_\lambda$. The arrival process of customers who are offered position $X_\lambda^{(n)}$ is Poisson with rate λ, thus, the queueing process is identical to that of an M/M/1/n system with arrival rate λ.

We observe that

$$\Pr\big(Q^{(n)} = k\big) = \int_0^{\overline{\lambda}} q_\lambda \cdot \Pr\big(X_\lambda^{(n)} = k\big) dF(\lambda), \quad \forall k \in \{1, \dots, n\}. \tag{3.1}$$

With respect to this representation of $Q^{(n)}$ and the potential arrival rate $\overline{\lambda}$, we reintroduce n_e, n_o and n_m as defined in (2.3). By Proposition 2.6, $X_\lambda^{(n)} \leq X_\lambda^{(n+1)} \leq X_\lambda^{(n)} + 1$ for every $\lambda \in [0, \overline{\lambda}]$, therefore $Q^{(n)} \leq Q^{(n+1)} \leq Q^{(n)} + 1$. This satisfies Assumption (A-ii) and by Corollary 2.12 we have that $n_m \leq n_o \leq n_e$. This explanation, in fact, applies to every concave decreasing sequence $\{u(k)\}_{k=1}^{\infty}$.

3.2.3 Heterogeneous Servers

Suppose now that servers are heterogeneous. We assume that waiting costs are incurred only when customers wait in the queue, but not in service[2]. Consider a customer who joins the queue in position k (i.e., upon arrival she observes $s + k - 1$ customers in the system in total). Her waiting time in the queue (excluding service), W_k, is Erlang distributed with shape parameter k and rate $\sum_{i=1}^{s} \mu_i$. Therefore, by Corollary 3.2, $\mathrm{E}(c(W_k))$ is convex as a function of k, hence, $u(k) = R - \mathrm{E}(c(W_k))$ is non-increasing and concave. From Corollary 2.12 we have that $n_m \leq n_o \leq n_e$.

[2] In general, customers paying for their own service times may violate Assumption (A-i): Depending on the service rates, customers may favor waiting in the queue to joining a free (slow) server.

3.3 Tandem Network Queues

The model of strategic customers in a network of queues (nodes) in tandem was first considered by D'Auria and Kanta (2015) for a two-node network, and was later extended by Kim and Kim (2016) to $m \geq 2$ nodes. To spare the reader additional technical complexity we present here a simplified version of the model in Kim and Kim (2016), yet we emphasize that more general results can be obtained using the complete analysis of both D'Auria and Kanta (2015) and Kim and Kim (2016).

Consider a tandem network consisting of m FCFS queues (nodes), each with an exponential service of rate μ. Customers arrive at the (first node of the) network following a Poisson process with rate λ. Upon arrival, each customer is informed of the *total* number of customers in the system (but not of the number at each specific node) and chooses to join the system or balk. The cost for a customer is $C > 0$ per unit time spent in the system [3]. After completing service at all nodes a customer receives a reward R.

Assuming customers follow some threshold n, it is shown in D'Auria and Kanta (2015) for $m = 2$ and later in Kim and Kim (2016) for $m > 2$, that, given the total number of customers in the system, the distribution of the number of customers within each node is independent of n. Therefore, so is the expected utility upon joining each position. Specifically, upon joining position $k \geq 1$ (i.e, when there are $k-1$ customers in the system in total), the expected utility, $u(k)$ is given by (see Kim and Kim (2016) Theorem 1)

$$u(k) = R - C\frac{k(k-1+m)}{\mu(k-1)},$$

which is strategy independent and monotone decreasing and concave in k. Moreover, all customers' service requirements (within each node, including those currently being served) are i.i.d., therefore by Proposition 2.6 Assumption (A-ii) holds. By Corollary 2.12, we immediately deduce that $n_m \leq n_o \leq n_e$.

3.4 Geo/Geo/1 Queue with Interupptions

Analyzing strategic behavior of *secondary users* (customers) in a Cognitive-Radio Network, Li and Han (2011) consider the following model: In discrete time epochs, customers' inter arrival times to a single-server FCFS station are Geo(λ)-distributed and service requirements are Geo(μ). When processing a job, interruptions (breakdowns)

[3] Both D'Auria and Kanta (2015) and Kim and Kim (2016) analyze the system with possibly different time cost and service rates at each node. Nevertheless, we note that in such a general setup, $u(k)$ may increase in k (see an example in D'Auria and Kanta (2015)§5).

may occur within random Geo(p)-distributed amount of time, during which the server halts for a random Geo(q)-distributed period, $\lambda, \mu, p, q \in (0,1)$. Interrupted services continue from the point they are stopped[4]. Customers can only join when the server is operating (that is, not during interruptions)[5]. Observing the queue length upon arrival (assuming the server is operating), they choose between joining or balking. A customer receives $R > 0$ upon service completion, and pays $C > 0$ per unit time waiting in the queue, but does not incur costs when being served, including time spent during an interruption of her own service [6]. For a customer joining position k, Li and Han show that

$$u(k) = R - \frac{C(k-1)}{\mu}\left(1 + \frac{p}{q}\right),$$

which is linear in k. Since the residual service requirements of all customers in the queue are i.i.d and memoryless, by Proposition 2.6 Assumption (A-ii) holds, and by Corollary 2.12 we obtain Naor's inequality.

3.5 The Abandonment Model

This following model, motivated by applications of order-driven markets, is similar to a one presented by Garman (1976). It is analyzed as a one dimensional birth-death process with state dependent transition rates (specifically, an M/M/1+M). Unlike the previous examples where customer utility was concave, here it turns out to be *convex* in their joined position. However, we can still apply Proposition 2.10 to show that Naor's inequality holds.

Consider customers who arrive, following a Poisson process with rate λ, at a single-server FCFS system with exponential service rate μ. Each customer may abandon the system within some amount of time, unknown upon arrival, which is an independent exponential random variable with rate θ. A customer can leave the system either by service completion or by abandonment, whichever comes first. We emphasize that abandonment occurs as a result of exogenous circumstances and *not* by customers' choice. Each customer incurs a fixed joining effort d upon joining[7]. If she completes service before

[4] Is is assumed that in a given epoch there may be several events of different types, but the possibility of two events of the same type is excluded.

[5] This assumption is not explicitly stated in Li and Han (2011) but is inferred by the proof in Li and Han (2011)§Appendix A. We note that relaxing this assumption, customers' utility depends on the probability that the server is operating conditioned on the number in the system, which in turn, depends on customers joining strategy. This case is similar to the continuous-time version partially-observable model of Economou and Kanta (2008b), but is inconsistent with the analysis in Li and Han (2011).

[6] This assumption in fact can be relaxed, but we adopt it for the sake of consistency with Li and Han (2011).

[7] d need not be positive.

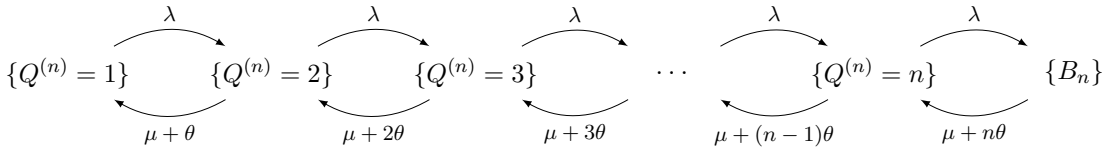


FIGURE 3.1: The Markov chain describing the queueing process of the Abandonment Model with threshold n.

abandonment she receives $R > 0$, and if she abandons she receives 0. For simplicity, assume that a customer may abandon the system at any time, even during her own service. The Markov Chain underlying the process is depicted in Fig. 3.1. Upon arrival, each customer observes the offered position and chooses whether to join or balk. Consider a customer who joins the system in position k, i.e., there are $k-1$ customers ahead of her. The probability that she will not abandon before any of these $k-1$ customers leaves is $1-\frac{\theta}{(\mu+k\theta)}$. When there are $k-2$ such customers, this probability becomes $1-\frac{\theta}{(\mu+(k-1)\theta)}$ and so forth. Thus, the probability that this customer will eventually complete service is given by

$$\frac{\mu+(k-1)\theta}{\mu+k\theta}\cdot\frac{\mu+(k-2)\theta}{\mu+(k-1)\theta}\cdot\ \ldots\ \cdot\frac{\mu}{\mu+\theta}=\frac{\mu}{\mu+k\theta}. \tag{3.2}$$

Let $u(k)$ denote the total expected utility of that customer. Equation (3.2) implies

$$u(k)=\frac{\mu}{\mu+k\theta}R-d. \tag{3.3}$$

Note that $u(k)$ is monotone decreasing and strictly convex.

At every moment in time, the residual service times of all customers are i.i.d, and so are each one's residual time to abandonment. Thus, Assumptions (A-i) and (A-ii) hold in the underlying model.

Corollary 3.4. *In the abandonment model described above,* $n_m \le n_o \le n_e$.

The proof of Corollary 3.4 is in Appendix B.2. In this proof we verify algebraically that the conditions of Proposition 2.10 are satisfied, using the algebraic expansions derived in Lemma 2.13.

3.6 The Server Catastrophe Model

The model discussed below was studied by Boudali and Economou (2012)[8]. What distinguishes this examples from all of our other examples is that here customers do

[8] Boudali and Economou (2012) studied both an observable and an unobservable versions of the model.

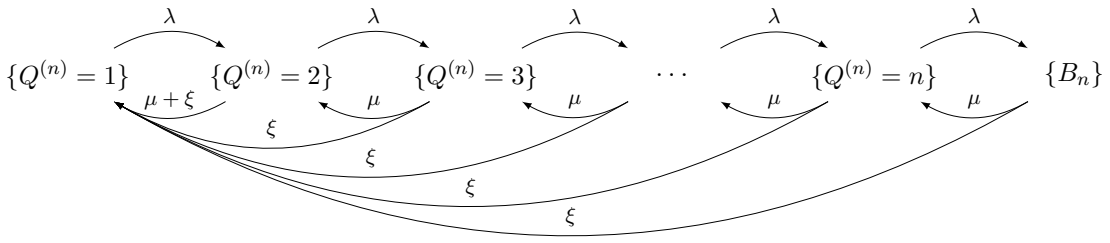


FIGURE 3.2: The Markov chain describing the queueing process of the Server Catastrophe Model with threshold n.

not necessarily leave the system one at a time, but may depart in bulks. Moreover, the expected utility as a function of position is convex, yet using Corollary A.5 and Lemmas 2.14 and 2.13 we prove Naor's inequality for this model.

Consider a single-server FCFS queue with independent exponential service and inter-arrival times with rates μ and λ respectively. The server is subject to *catastrophes* which occur according to an independent Poisson process with rate ξ [9]. When a catastrophe occurs, all customers in the system leave immediately without service. We note that since there is a non-zero transition rate from any possible state to an empty system, the system is not a birth-death process (see Fig. 3.2).

Following the notation in Boudali and Economou (2012), each joining customer may depart from the system either by completing service and receiving a reward R_s, or by the occurrence of a catastrophe, receiving a catastrophe compensation R_f. Customers incur waiting cost C per unit time waiting in the system. Upon arrival, each customer observes the queue length and chooses either to join or balk. Boudali and Economou show that the expected utility for a customer joining position k is given by

$$u(k) = V \cdot \left(\frac{\mu}{\mu+\xi}\right)^k + D, \tag{3.4}$$

where $V = R_s - R_f + \frac{C}{\xi} > 0$ and $D = R_f - \frac{C}{\xi}$. Since $0 < \mu < \mu + \xi$, $u(k)$ is a decreasing *convex* sequence.

At every moment, the residual service times of all customers are i.i.d, therefore, changing the queueing regime does not affect the stationary distribution of the system. In particular, we can, as in the proof of Proposition 2.6, interpret the event A_n as the

[9] In Boudali and Economou (2012) it is assumed that after a catastrophe, the time for the server to recover is exponential, and within this recovery period customers are not accepted into the system. Since the recovery time is irrelevant for the decision making process we ignore it and assume the recovery after catastrophe is immediate.

existence of the standby customer[10], thus, Assumption (A-ii) holds. For simplicity, we assume $\lambda \leq \mu$[11].

Lemma 3.5. *In the Server Catastrophe model, when* $\lambda \leq \mu$*,*

(i) $\Pr(B_n) \leq \Pr\big(Q^{(n)} = k\big)$*, and*

(ii) $\Pr\big(Q^{(n+1)} = k\big) \leq \Pr\big(Q^{(n)} = k\big)$*,*

for all $n \in \mathbb{N}$ *and* $k \leq n+1$*.*

The proof of Lemma 3.5 is in Appendix B.3.

Lemma 3.6. *In the Server Catastrophe model, when* $\lambda \leq \mu$*,*

$$\Pr(A_n) \leq \left(\frac{\mu}{\mu+\xi}\right)^{n-1}.$$

The proof of Lemma 3.6 is in Appendix B.4.

From Lemmas 3.5 and 2.14 we immediately obtain $\Pr(A_n \mid J_n) \leq \Pr(A_n)$ and from Equation (3.4), $u'(k) = V \cdot \left(\frac{\mu}{\mu+\xi} - 1\right)\left(\frac{\mu}{\mu+\xi}\right)^k$. Using Lemma 3.6,

$$\Pr(A_n \mid J_n) \leq \Pr(A_n) \leq \left(\frac{\mu}{\mu+\xi}\right)^{n-1} = \frac{V \cdot \left(\frac{\mu}{\mu+\xi} - 1\right)\left(\frac{\mu}{\mu+\xi}\right)^n}{V \cdot \left(\frac{\mu}{\mu+\xi} - 1\right)\left(\frac{\mu}{\mu+\xi}\right)} = \frac{u'(n)}{u'(1)},$$

for all $n \in \mathbb{N}$. Therefore, by Corollary 2.11 we conclude that $n_m \leq n_o \leq n_e$.

3.7 Many Queues System

In Chapter 2 and in the examples in §3.2-§3.6 we refer to a customer's joining position naturally as her physical waiting slot, meaning that a customer joining in position k has to wait for $k-1$ customers to begin service. In fact, our modeling in Chapter 2 allows for more abstract interpretations of the term joining position. The following example demonstrates that the joining position, by its most general interpretation, simply represents an information signal a customer receives upon arrival, and does not necessarily determine the number of customers that a joining customer has to wait for before being

[10] Like every other customer, this standby customer can either leave when completing service or when a catastrophe occurs

[11] Similar results can be obtained for more general cases, but, require more prudent analysis.

served. In the spirit of the discussion in §2.5.2, this example also shows that strategy-independent utility need not imply that customers are served FCFS with respect to their arrival order.

Suppose that customers arrive according to a Poisson process with rate λ at a system composed of $m \in \mathbb{N}$ symmetric FCFS queues, each of which is served by a separate server with exponential service time of rate μ. Upon arrival, a customer is informed of the total number of customers in the system. Thus, her offered position represents the total number of customers in the system a moment after arrival if she joins. When joining, the customer is sent to one of the queues at random, with probability $\frac{1}{m}$ to each queue[12]. The reward from service is $R > 0$ and the cost per unit time waiting is $C > 0$. We note that even though each queue is served FCFS, the overall service regime is not FCFS, as it is possible that a customer will begin service before an earlier-arriving customer, provided they have drawn different queues.

By symmetry and the PASTA property, when the total number of customers is k, the expected utility for a joining customer is given by $u(k) = R - \frac{Ck}{m\mu}$, which is linear in k. Since at any instant, the residual service of all customers, including those currently being served, are i.i.d, by Proposition 2.6, Assumption (A-ii) holds. We therefore conclude, by Corollary 2.12, that $n_m \leq n_o \leq n_e$.

3.8 Compartmenented Waiting Space

Unlike §3.7, in the following example again we consider customers offered positions as their physical waiting space. However, this example is different from §3.2-§3.6 in the sense that several customers waiting together in the queue may share the same position.

The model is discussed in a paper by Economou and Kanta (2008a). Similar to Naor (1969), the queue is M/M/1, FCFS, with arrival and service rates λ and μ respectively. Customers receive a reward R from service completion and incur cost of C per unit time in the system. However, in Economou and Kanta (2008a), the waiting space in the queue is partitioned into compartments of fixed capacity of A customers each. An arriving customer observes her offered *compartment number*[13], that is, if the total number of customers prior to her arrival is k, she observes $\lfloor k/A \rfloor + 1$. It is assumed that customers follow a pure threshold strategy, i.e., they join if and only if the offered compartment number is no greater than a threshold n. Given a threshold n, we consider a customer's

[12] Because of the random picking of the queue, concerning her expected waiting time, it is irrelevant whether a customer can observe only the total number of customers in the system or the exact number at each queue.

[13] Economou and Kanta (2008a) studied both the case of observable compartment numbers, and the case that only the number of customers in the last compartment is observable.

stationary offered position $Q^{(n)}$ as the offered compartment number, and define $S^{(n)}$ and $M^{(n)}$ accordingly as in Equations (2.1) and (2.2). Thus, n_m, n_o and n_e represent the thresholds on the number of *compartments* chosen by the monopolist, the social planner and the individuals, respectively. Economou and Kanta (2008a) prove algebraically that $n_m \leq n_o \leq n_e$ solving for the stationary offered position distribution. We provide below an alternative proof applying Corollary 2.12:

Let $X^{(n)}$ be the stationary number of customers in an M/M/1/n system. By Proposition 2.6, $X^{(n)} \leq_{\text{st}} X^{(n+1)} \leq_{\text{st}} X^{(n)} + 1$, thus, a simple induction argument gives

$$X^{(n\cdot A)} \leq_{\text{st}} X^{(n+1)\cdot A} \leq_{\text{st}} X^{(n\cdot A)} + A.$$

From the PASTA property, and the definition of the offered position, $Q^{(n)} \sim \lfloor X^{(n\cdot A)}/A \rfloor + 1$. Therefore $Q^{(n)} \leq_{\text{st}} Q^{(n+1)} \leq_{\text{st}} Q^{(n)} + 1$, and Assumption (A-ii) holds.

Let $\rho = \frac{\lambda}{\mu}$. It is shown in Economou and Kanta (2008a) that a customer's utility from joining position k is given by:

$$u(k) = \begin{cases} R - \frac{C}{\mu}\left(k \cdot A - \frac{A}{1-\rho^A} + \frac{1}{1-\rho}\right) & \text{if } \rho \neq 1, \\ R - \frac{C}{\mu}\left(k \cdot A - \frac{A-1}{2}\right) & \text{if } \rho = 1. \end{cases}$$

Thus $u(k)$ satisfies Assumption (A-i). Since it is also linear, by Corollary 2.12, $n_m \leq n_o \leq n_e$.

3.9 The FCFS-Orbit Constant-Retrial Queue

This example refers to a model studied by Economou and Kanta (2011)[14]. As in §3.7, the service regime as a whole is not FCFS here, yet the utility from joining each position is independent of customers strategy. The authors of Economou and Kanta (2011) prove Naor's inequality, by explicitly solving for the stationary distribution of the offered position in the system. We provide here a new proof of Naor's inequality using the machinery developed in Chapter 2.

Consider a single server with an FCFS *orbit queue* and constant retrial rate: Service demand is exponential with rate μ. Potential customers arrive following a Possion process with rate λ. Upon arrival, if a customer finds an idle server, she is immediately admitted into service. If she finds it busy, then she chooses between joining an orbit queue or leaving. Unlike many orbit models, where each customer in orbit employs an independent retrial process, here the orbit queue is FCFS: Customers in orbit are ordered according to their arrivals, and only the *first* customer in orbit generates a retrial

[14] Economou and Kanta (2011) studied both an observable and an unobservable versions of the model.

process which is Poisson with rate α. Hence, the overall process of trials made to the server, is Poisson with rate $\lambda+\alpha$ given the orbit is nonempty, and with rate λ otherwise.

It is assumed that an arriving customer observes the server's state and the number of customers in orbit. As explained, the number in orbit is only relevant to her decision when arriving at a busy server, otherwise she begins service instantaneously. Customers who join the system receive a reward R for completing service and incur waiting cost C per unit time spent (both in service and orbit).

Depending on the server's occupancy, an arriving customer may begin service before customers in orbit. Thus, we refer to the offered position as 1 when the server is idle, 2 when a customer is offered the head of the orbit, 3 when offered to be second in orbit and so forth.

Economou and Kanta (2011) show that customer expected utility for joining position k is given by

$$u(k) = R - C\left(\frac{\lambda+\alpha+\mu}{\mu\alpha}(k-1) + \frac{1}{\mu}\right). \tag{3.5}$$

Although customer utility is affected by future arrivals (and varies with λ), it is strategy independent. It can be seen from (3.5) that $u(k)$ is decreasing, i.e., Assumption (A-i) holds. Since the service requirements of all customers are i.i.d exponential variables, by Proposition 2.6, Assumption (A-ii) holds as well[15]. Finally, since $u(k)$ is linear (and therefore concave), by Corollary 2.12, $n_m \leq n_o \leq n_e$.

Zhang et al. (2014) studied an extension of this model coping with server breakdowns, that occur within an exponential time of rate ξ, and last for an exponential duration of rate η. Breakdowns can only occur when the server is busy, and during breakdown customers are not admitted to the system. There, it is shown that

$$u(k) = R - C\left(\frac{\xi+\eta}{\mu\eta} + \frac{(\lambda+\alpha)\cdot(\xi+\eta)+\mu\eta}{\alpha\mu\eta}k\right),$$

which, as before, is a linear decreasing function. In Zhang et al. (2014), Naor's inequality is observed numerically. Indeed, Corollary 2.12 applies with no additional effort also for the case of breakdowns and repairs studied in Zhang et al. (2014).

[15]The coupling construction used here is identical to that in the proof of Proposition 2.6, assuming that a customer joining the system in position $n+1$ is a standby customer. Here, the standby customer, when preempted during service, is going back to the orbit, while in the orbit she is assigned the lowest priority.

3.10 M/G/1 with Concave Utility

Altman and Hassin (2002) study strategic customer behavior in an observable FCFS M/G/1 system with constant reward R and waiting cost C. They note that in such systems, customer expected utility may not be strategy independent: To calculate the expected utility from joining, one has to assess the residual time of the current service given the offered position. Assuming that customers can also employ mixed joining strategies, this assessment of the residual service may depend on customers' joining strategy, calling for a game-theoretic analysis. Furthermore, this expected utility is not necessarily monotone in the offered position, thus, violating Assumption (A-i). Boxma (1984) and Kerner (2008) discuss ways to compute the mean sojourn time of a customer given her joining position. Kerner (2011) considers a model similar to Altman and Hassin (2002) and specifies the equilibrium structure that emerges for different families of service distributions. In particular, conditioned on joining position k, the residual service time does not depend on the strategy of customers joining positions higher than k. This means that if customers are *assumed* to follow threshold joining, then customer expected utility is strategy independent. The reason, as explained by Kerner (2011), is the following: Given the server is busy and $k \geq 0$ customers are awaiting, all those customers who had arrived from the moment the current service started clearly observed less than k customers in the system and therefore joined. Yet again, we emphasize that an equilibrium in threshold strategies may not exist in an M/G/1 model with the utility structure considered in Altman and Hassin (2002) and Kerner (2011).

Regardless of the structure of utility, we make the following observation:

Lemma 3.7. *Assumption (A-ii) holds in an M/G/1 (non-preemptive) system.*

The proof of Lemma 3.7 is in Appendix B.5.

To comply with our model assumptions in Chapter 2, we suppose that $u(k)$ is strategy independent and decreasing. By Corollary 2.12 we get, as an immediate result of 3.7,

Corollary 3.8. *In the observable M/G/1 with concave decreasing (and strategy independent) $u(k)$, $n_m \leq n_o \leq n_e$.*

A special case of this model is considered by Sun et al. (2018), who study an observable M/G/1 queue, where customers posses partial information (which is common knowledge) about the service distribution. Given this partial information and the joining position k, a customer forms a belief about her waiting time, T_k, which by assumption is a sum of k independent random variables, $T_k = \tilde{S}_1 + S_2 + \cdots + S_k$, where $S_2, \ldots, S_k$ are identically

distributed[16]. The distributions for $\tilde{S}_1$ and for $S_2, \ldots, S_k$ are assessed by each customer based on the (common) partial information. Then, for given parameters R,C and $m \geq 0$, the expected utility $u(k)$ depends on the $(m+1)$st moment of T_k. Specifically, it takes the form $u(k) = R - C\mathrm{E}\left(T_k^{m+1}\right)$, which, by Lemma 3.1, is concave decreasing in k. Therefore by Corollary 3.8, $n_m \leq n_o \leq n_e$. This was verified numerically for several sets of parameters of choice in Sun et al. (2018).

3.11 Examples for $n_o < n_m$

In this subsection we discuss cases where the inequality $n_m \leq n_o$ does not hold. Particularly, the model in §3.11.1 satisfies assumptions (A-i) and (A-ii), emphasizing why these two assumptions alone do not guarantee Naor's inequality. The model in §3.11.2 assumes linear decreasing utility, thereby satisfying Assumption (A-i) and concavity (as in Corollary 2.12), but violates Assumption (A-ii).

3.11.1 Assuming (A-i) and (A-ii)

To simplify the analysis, we will assume an M/M/1, FCFS queueing model (thus, satisfying Assumption (A-ii)) with identical arrival and service rates, $\lambda = \mu$. With respect to the joining-position distribution, we deliberately choose the form of $u(k)$ to induce $n_o < n_m$. To this aim, $u(k)$ must be non-concave, otherwise by Corollary 2.12 Naor's inequality holds. Specifically, we will assume here, for some constant R, that $u(k) = \frac{1}{k} + R$ which is decreasing (hence satisfying (A-i)) and (strongly) convex.

Given threshold n, the offered position is uniformly distributed, $Q^{(n)} \sim \mathrm{U}\{1, \ldots, n+1\}$. Suppose that $R = \frac{3}{2}$. By (2.1) and (2.2) we obtain

$$\mathrm{E}\big(M^{(n)}\big) = \Pr(J_n) \cdot u(n) = \frac{n}{n+1} u(n) = \frac{n}{n+1} \cdot \left(\frac{1}{n} + \frac{3}{2}\right) = 1 + \frac{1}{2} \cdot \frac{n}{n+1},$$

and

$$\mathrm{E}\big(S^{(n)}\big) = \frac{1}{n+1} \sum_{k=1}^{n} u(k) = \frac{1}{n+1} \left(\frac{3n}{2} + \sum_{k=1}^{n} \frac{1}{k}\right).$$

$\mathrm{E}\big(M^{(n)}\big)$ is monotone increasing[17] in n and therefore $n_m = \infty$. Considering $\mathrm{E}\big(S^{(n)}\big)$, by Knudsen (1972)§4 Theorem 1, we have that $\mathrm{E}\big(S^{(n)}\big)$ is unimodal. Figure 3.3 depicts

[16] For unspecified reasons, Sun et al. (2018) assume that the distribution of $\tilde{S}_1$ does not depend on the observed queue length, k (which also implies that $T_1 = \tilde{S}_1$).

[17] In this example, $u(n) \geq 0$ for all n, implying $n_e = \infty$. A similar example assuming $u(k) = \frac{1}{k} + R - C \cdot k$, for sufficiently small $C > 0$, will result in the same qualitative observation $n_o < n_m$, but also $n_e = \lfloor (R + \sqrt{4C + R^2})/(2C) \rfloor$ therefore $n_m \leq n_e < \infty$.

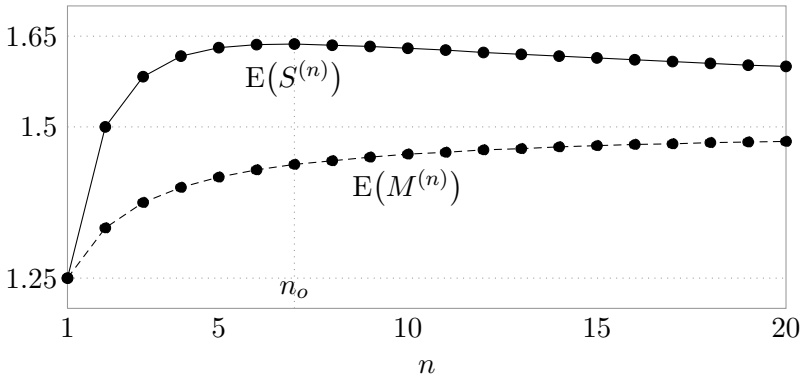


FIGURE 3.3: $\mathrm{E}(S^{(n)})$ and $\mathrm{E}(M^{(n)})$ as functions of n, in an M/M/1 with $\lambda = \mu = 1$ and $u(k) = \frac{1}{k} + \frac{3}{2}$.

$\mathrm{E}(S^{(n)})$ and $\mathrm{E}(M^{(n)})$ as functions of n. Note that $n = 7$ is a local (hence, also global) maximum of $\mathrm{E}(S^{(n)})$, therefore $n_o = 7 < n_m = \infty$.

Inspired by models of *benchmark effects*[18], we provide an interpretation for $u(k)$: Each service has a quality which is drawn independently at random from some continuous distribution. At the end of her service, if a customer's service quality was the highest among all the services completed while she was in the system, then she considers it an "extraordinary" service, and values it as $R + 1$. Otherwise, her valuation is R. In particular, the service valuation of a customer who arrives at an empty station is $R+1$. Given a customer joining position k, the probability that her service would be of the highest quality among all k services is $\frac{1}{k}$. Thus, $u(k) = \frac{1}{k} + R$.

An alternative interpretation for $u(k)$ is that customer valuation of service deteriorates over time, taking the form $R + \frac{1}{\mu t}$ for t the time spent in the system. Assume that each customer, after joining position k and waiting for k service completions (including her own) has to wait for an additional independent Exp(μ)-distributed lay-off. Then her total time spent in the system is Erlang$(k+1, \mu)$ distributed, therefore

$$u(k) = R + \frac{1}{\mu}\int_{t=0}^{\infty} \frac{1}{t} \cdot \frac{\mu^{k+1}t^k}{k!} e^{-\mu t} dt = R + \frac{1}{k}\int_{t=0}^{\infty} \frac{\mu^k t^{k-1}}{(k-1)!} e^{-\mu t} dt = \frac{1}{k} + R,$$

as desired.

We note that in the Abandonment Model in §3.5, similarly, $u(k)$ is decreasing at the rate of $\frac{1}{k}$. In particular, considering Equation (3.3), for $\theta = 1$, $d = -\frac{3}{2}$ and $R = \frac{1}{\mu}$, taking $\mu \to 0$ will result in the same utility function, $\frac{3}{2} + \frac{1}{k}$ as here. Yet, we show in §3.5, that for every choice of parameters in the Abandoment Model, $n_m \leq n_o$, in contrast to the result in this section. The essential difference between these two examples is that, due

[18] These are models where customers compare their own service quality with service quality of others, for more information, see Hassin (2016)§4.3.

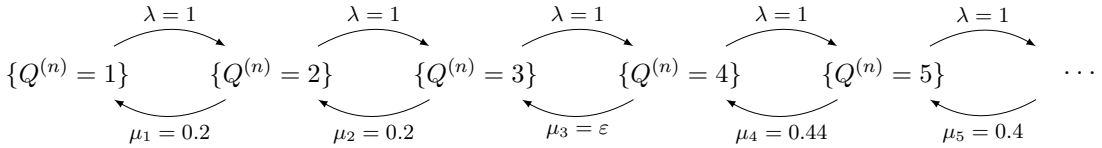


FIGURE 3.4: The Markov chain describing the queueing process of the state-dependent service rates model.

to the abandonment process (which is state-dependent), the level of congestion in the Abandoment Model increases relatively slowly as the threshold increases. However, in the M/M/1 example presented here, the congestion level is more 'sensitive' to changes in the threshold. In line with the intuition provided for Proposition 2.10, the conditions of the proposition hold for the Abandonment Model, but not for the M/M/1 model in this section.

3.11.2 Relaxing Assumption (A-ii)

In order to clarify why Assumption (A-ii) is essential to our discussion we demonstrate how its relaxation affects Naor's inequality. In light of the discussion in §2.5.2, we emphasize that Assumption (A-ii) is natural in many common queueing models, thus designing a model violating it calls for cooking up a relatively pathological example. Specifically, the example below is tailored such that n_o and n_m do not obey Naor's inequality, in spite of customer utility being linear decreasing (thus, satisfying Assumption (A-i)) and the queue being a birth-death process.

This example studies a special case of an M/M/1 system with non-monotone, state-dependent service rates. A closely related model with non-decreasing rates is discussed by Burnetas and Dimitrakopoulos (2018), and the equilibrium behavior is derived. Here we assume a different utility structure, linearly decreasing in the joining position, and strategy independent.

Consider an M/M/1 system with arrival rate λ and state-dependent service rates, such that when $k \geq 1$ customers are present (thus, the offered position is $k+1$) the service rate is μ_k. The utility is linear in the joining position, $\{u(k)\}_{k=1}^{\infty} = \{10, 8.5, 7, 5.5, \dots\}$. Of course, if the service rates are chosen such that the model satisfies Assumption (A-ii), by Corollary 2.12, we immediately obtain Naor's inequality.

Suppose that $\lambda = 1$, $\mu_1 = \mu_2 = 0.2$, $\mu_3 = \varepsilon$, $\mu_4 = 0.44$, and $\mu_k = 0.4$ for all $k \geq 5$ (see Fig. 3.4). Note that for $n = 1$ and $n = 2$ the model coincides with Naor (1969), and

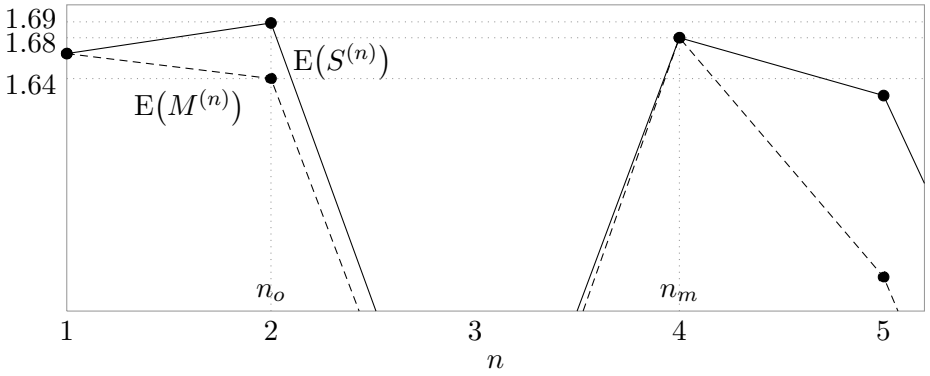


FIGURE 3.5: $\mathrm{E}(S^{(n)})$ and $\mathrm{E}(M^{(n)})$ as functions of n, in the state-dependent service rates model for sufficiently small $\varepsilon > 0$.

standard algebraic calculations yield

$$\mathrm{E}(S^{(1)}) = \mathrm{E}(M^{(1)}) = \frac{u(1)}{1+\frac{\lambda}{\mu_1}} = 1.667,$$

$$\mathrm{E}(S^{(2)}) = \frac{u(1)}{1+\frac{\lambda}{\mu_1}+\frac{\lambda^2}{\mu_1\mu_2}} + \frac{u(2)\frac{\lambda}{\mu_1}}{1+\frac{\lambda}{\mu_1}+\frac{\lambda^2}{\mu_1\mu_2}} = 1.694, \quad \mathrm{E}(M^{(2)}) = \frac{u(2)\left(1+\frac{\lambda}{\mu_1}\right)}{1+\frac{\lambda}{\mu_1}+\frac{\lambda^2}{\mu_1\mu_2}} = 1.645.$$

When $\varepsilon = 0$, the limiting distribution of $Q^{(3)}$ implies $\Pr(Q^{(3)} = 4) = \Pr(B_3) = 1$. Clearly, $Q^{(2)} + 1 \leq_{\mathrm{st}} Q^{(3)}$ (in the strong sense), violating Assumption (A-ii) for $n = 2$. Thus, as ε approaches 0, both $\mathrm{E}(S^{(3)})$ and $\mathrm{E}(M^{(3)})$ approach 0 (the system with $n = 3$, when stationary, reaches an absorbing state where all customers balk). Moreover, if we take $\varepsilon = 0$ and $n = 4$, then $\Pr(Q^{(4)} \in \{4, 5\}) = 1$ and it can be easily verified that

$$\mathrm{E}(S^{(4)}) = \mathrm{E}(M^{(4)}) = \frac{u(4)}{1+\frac{\lambda}{\mu_4}} = 1.681 \in \left(\mathrm{E}(M^{(2)}), \mathrm{E}(S^{(2)})\right),$$

with both $\mathrm{E}(S^{(n)})$ and $\mathrm{E}(M^{(n)})$ decreasing for $n \geq 4$. Therefore, there exists an ε sufficiently small such that $\mathrm{E}(M^{(4)}) > \mathrm{E}(M^{(1)}) > \mathrm{E}(M^{(2)}) > \mathrm{E}(M^{(3)})$, and $\mathrm{E}(S^{(2)}) > \mathrm{E}(S^{(4)})$, implying that $n_o = 2 < n_m = 4$ (see Fig. 3.5). We note that when $\varepsilon = 0$ and the chosen threshold is $n = 3 + m$, the system can be viewed as a separate system with threshold m, with the sequence of utility being $\{5.5, 4, 2.5, \dots\}$. Following this perspective, each of the two thresholds $n = 2$ and $n = 4$ induces a different queueing system unrelated to the other, and, by choosing the right parameters μ_1, μ_2 and μ_4, any ordering of the terms n_o and n_m is achievable.

3.12 Concluding Remarks

Traditionally, the individually optimal threshold n_e is referred to as the equilibrium threshold. Yet when $u(k)$ is independent of customers' strategy as assumed here, the threshold strategy n_e is a dominant strategy for every customer, so the game among customers is a degenerate one. Ample existing literature deals with models where the utility of a customer depends on both the offered position and the threshold. In such models the interaction between customers indeed brings about a game where customers form beliefs on the strategy of others before making their own decision. An example we refer to in §2.5.1 is given by Altman and Shimkin (1998) where customers with linear waiting costs join or balk from an observable egalitarian processor-sharing (EPS) system. There, customer utility is a function of both the offered position and the threshold, and decreases in both variables. An interesting future-research direction is to extend our results to such models.

PART II: Unobservable Queues

Chapter 4

Strategic Customer Behavior in Cognitive Radio Networks

4.1 Background and Motivation

It is sometimes the case that customers choose between queues of different types or discipline (see Hassin (2016)§8.2 for a survey). In particular, we refer to situations that involve choosing between a blocking subsystem and a shared subsystem of an unlimited capacity. The information with which customers are provided plays a crucial role in the decision making process, and occasionally, customers are willing to allocate their resources (money, time or energy) for acquiring information (as in Xu and Hajek (2013), Roet-Green and Hassin (2014), for more information see Hassin (2016)§3.4).

The information structure of the model presented here is an instance of the *unobservable model*, in which customers are not informed of the queue length upon arrival. The concept of the unobservable M/M/1 queue was first introduced by Littlechild (1974) and by Edelson and Hildebrand (1975), and is covered in detail by Hassin and Haviv (2003). Roet-Green and Hassin (2014) study a variation of the unobservable model where customers are offered to purchase information about the queue length before deciding whether to join or balk. They show that customers who purchase information about the queue length induce positive externalities on others. Another related model is the *supermarket game* studied by Xu and Hajek (2013) in which customers inspect one or more queues out of N unobservable queues, where the number of queues to inspect, k, is the customer's decision variable, and the inspection cost is linear in k. There, *mean field* approximation is applied for specifying a symmetric equilibrium when $N \to \infty$. It is shown, that in the mean field model in Xu and Hajek (2013), inspection induce positive externalities, as holds for the model of Roet-Green and Hassin (2014). When N is finite,

this observation does not hold in general. The authors give an example with $N = 2$ where customers can choose between joining one of the queues at random or inspecting both and joining the shortest, and show that inspection may induce negative externalities on future customers. Yet, this fact alone is not enough to determine whether the proportion of inspecting customers in equilibrium would be more or less then optimal. In this work we show that not only purchasing information can sometimes impose negative externalities on others, but also that sometimes purchasing information decreases the social welfare even when the individual may find it profitable for him/her to do so.

A prominent application motivating the research on unobservable queues and/or information acquisition comes from the framework of *cognitive radio networks* (CRNs), which are naturally associated with unobservable queueing systems (as in Do et al. (2012), Habachi and Hayel (2012), Hayel et al. (2014) and Jagannathan et al. (2012)). Do et al. (2012) and Jagannathan et al. (2012) study a generalization of the unobservable M/M/1 queue that models customer behavior in a CRN with users of two different types. Corresponding to Do et al. (2012) and Jagannathan et al. (2012), Li and Han (2011) study an observable variation of a queue with server breakdowns (in which customers make their decisions based on the observed queue length). A type of CRNs which, according to a survey paper by Haykin (2005), is of a particular interest is the sensing-based networks. In sensing-based CRNs, users are endowed with the ability to sense the channel, i.e., listen to the channel and detect whether or not it is free for transmission. Using this ability, users can allocate their transmit power while maximizing their own benefit. As the act of sensing can be expensive (in terms of energy consumption, time, etc.), some users might waive sensing and join an unlicensed, freely-shared channel. Habachi and Hayel (2012) investigate the decision process of users choosing between sensing and not sensing in a system consisting of two classes of customers, primary and secondary. Upon arriving to the system in Habachi and Hayel (2012), each customer chooses between joining a shared M/M/1 or applying (i.e. "sensing") for a free server in an M/M/s/s loss system.

Another related application is of *last mile delivery services*, where customers choose between two service regimes: door-to-door delivery or self pickup. Driven by this problem, Hayel et al. (2014) consider a decision problem of customers choosing between an unobservable M/M/s/s and an unobservable M/D/1 queue. Other fields of practice include cloud computing services and the like.

Habachi and Hayel (2012) and Hayel et al. (2014) assume that sensing customers rejected by the blocking system leave never to return. This approach simplifies the analysis as it permits the division of the system into two queuing subsystems with two independent Poisson arrival streams. However, a model that permits customers to leave the system

without being served raises concerns regarding the utility of service; Habachi and Hayel (2012) consider a problem of minimizing costs and neglect the reward, assuming that blocked customers skip both waiting and service and end up paying nothing. In this case, blocking becomes beneficial for the customers as it minimizes their total expense. Yet, in many queueing applications, blocking is considered by the customers an undesirable outcome, because service is associated with a reward. One can assume a fixed service reward and formulate the customers decision problem as a maximization problem, although customers then must be given the opportunity to balk from the system when they would otherwise be forced to bear a negative expected return. A natural question therefore would be how the decision process in Habachi and Hayel (2012) and Hayel et al. (2014) would be changed when allowing blocked customers to join the shared queue.

Presented in this chapter is a model closely related to the one in Habachi and Hayel (2012) and Hayel et al. (2014), but the assumptions made on the consequences of rejection make a significant difference. We study a network consisting of two servers, one queue, and homogeneous customers. One of the two servers in this model is a loss system (M/M/1/1); customers trying to enter the loss system (for the sake of avoiding congestion) are exposed to the risk of rejection. In addition, customers are charged for their act of sensing even when ended with rejection, though it is assumed to be instantaneous. As opposed to previously mentioned models, here we assume that sensing customers, upon encountering rejection, are redirected to the shared queue. Thus, the queue in the system contains both customers who initially choose to join it as well as blocked sensing customers. This assumption induces dependencies between the state of the loss system and the inter-arrival times of customers to the queue. More formally, the queue within this model functions as a modification of an M/M/1 queue in which the rate of arrivals is subject to Poisson alternations. Such Heterogeneous Arrivals Queues, were analyzed by Yechiali and Naor (1971), and the results lay the foundations of the queueing analysis in this work. For the sake of keeping the model simple and reducing the number of parameters, we assume that service times for both servers have identical rates. All our results can be easily extended to the case of different exponential rates for each server.

Our key results in this chapter are two: First we show that there exists a unique equilibrium strategy, which in general, but not always, is different from the socially-optimal strategy. Second, depending on the system parameters, customers in equilibrium may sense the loss system either in a higher, lower or in an equal rate to what the social planner would desire. This arises counter intuitively inasmuch as selfish behavior in most models always yields an excessive joining rate to the shared queue, due to negative joining externalities (see Edelson and Hildebrand (1975), Hassin and Haviv (2003)

and Littlechild (1974)). Yet, these cases which we call *over-sensing* are rare and their significance in serving any practical purpose is arguable.

We begin by presenting the mathematical model in §4.2. In §4.3 we derive stability conditions for the system. Afterwards, in §4.4 and §4.5 we find the equilibrium and the socially-optimal strategies and compare them using the concept of *price of anarchy*. Finally, in §4.6 we discuss possible future ways of generalizing the model for further investigation.

4.2 Model Description

We consider a system composed of two identical servers, S_Q and S_L (Q for queue, L for loss) and a single FCFS unobservable queue. Each service duration is exponential with rate μ. Customers are identical, and arrive at the system following a Poisson process with rate Λ. Upon arrival, a customer chooses one of two options: pay a *sensing price* and try to attain service in S_L, an action we shall call *sensing*, or join the shared queue and wait until accepted to service in S_Q, which will be named *joining*. Customers sensing S_L when the server is idle are immediately accepted to service, whilst sensing S_L when busy, they get rejected by S_L and redirected to the shared queue to wait their turn for service by S_Q.

Denote by p the probability that a customer senses S_L, and denote by $(X(t), Y(t))$ the state of the system at time t. $X(t) \in \{0, 1, 2, \ldots\}$ expresses the number of customers in the queue including the one in service in S_Q. $Y(t) \in \{0, 1\}$ expresses the state of S_L, where $Y(t) = 0$ means that S_L is idle at time t. Thus, the system can be described as a bi-dimensional Markov process with one bounded dimension. Let $P_{i,j}$ be the stationary probability that $X(t) = i$ and $Y(t) = j$ at some arbitrary moment t. In order to simplify the notation, when dealing with stationary probabilities we omit the use of t and denote the state by (X, Y), and $P_{i,j} = \Pr(X = i, Y = j)$. The stationary probabilities are computed through the following set of equations:

$$\begin{aligned} &\Lambda P_{0,0} - \mu P_{1,0} - \mu P_{0,1} = 0 \\ &(\mu + \Lambda) P_{0,1} - p\Lambda P_{0,0} - \mu P_{1,1} = 0 \end{aligned} \tag{4.1}$$

$$\forall n \in \{1, 2, \ldots\} : \begin{cases} (\mu + \Lambda) P_{n,0} - (1-p)\Lambda P_{n-1,0} - \mu P_{n+1,0} - \mu P_{n,1} = 0 \\ (2\mu + \Lambda) P_{n,1} - p\Lambda P_{n,0} - \Lambda P_{n-1,1} - \mu P_{n+1,1} = 0\,. \end{cases} \tag{4.2}$$

These equations lead to the following relationship:

$$\forall n \in \{0,1,2,\ldots\}: \; P_{n+1,0} + P_{n+1,1} = \frac{\Lambda}{\mu}((1-p)P_{n,0} + P_{n,1}) \,. \tag{4.3}$$

By the assumption that each individual chooses randomly and independently whether or not to sense, we deduce that subsystem S_L can be considered as an M/M/1/1 queue (Erlang's Loss Model) with arrival rate $p\Lambda$ and service rate μ. Let $\rho := \frac{\Lambda}{\mu}$; then the *loss-probability* of S_L, i.e., the probability that a customer sensing S_L will find it busy, is:

$$\Pr(Y=1) = 1 - \Pr(Y=0) = \sum_{i=0}^{\infty} P_{i,1} = \frac{p\Lambda}{\mu + p\Lambda} = \frac{p\rho}{1+p\rho}, \tag{4.4}$$

Summing the equations in (4.3) over $n = 0, 1, 2, \ldots$ and utilizing (4.4) we get:

$$P_{0,0} + P_{0,1} = 1 - \rho\left(1 - \frac{p}{1+p\rho}\right) \tag{4.5}$$

which indicates a linear relationship between $P_{0,0}$ and $P_{0,1}$. Subtracting the second equation from the first one in (4.2) and combining with (4.3) we have:

$$\forall n \in \{0,1,\ldots\}: \; \begin{cases} P_{n+1,0} = (1-p)\rho P_{n,0} + p\rho \sum_{i=0}^{n} P_{i,0} - \sum_{i=0}^{n} P_{i,1} \,, \\ P_{n+1,1} = \rho P_{n,1} + \sum_{i=0}^{n} P_{i,1} - p\rho \sum_{i=0}^{n} P_{i,0} \,. \end{cases} \tag{4.6}$$

It can be seen from (4.6) that, for given p and ρ, every value $P_{n+1,k}$,$k \in \{0,1\}$, is expressed as a linear combination of the values $\{P_{0,0}, P_{1,0}, \ldots, P_{n,0}, P_{0,1}, P_{1,1}, \ldots, P_{n,1}\}$. Hence, with (4.5), it follows by induction that for each $n \in \{0,1,\ldots\}$ and for each $k \in \{0,1\}$, $P_{n,k}$ is a linear function of $P_{0,0}$, as a countable sum of linear functions.

An accurate solution for (4.6) (including an analytic solution for $P_{0,0}$) is presented in a paper by Yechiali and Naor (1971), but in an implicit form, since the solution makes use of Cardano's formula. The method in Yechiali and Naor (1971) can be utilized in evaluating the quantities $\mathrm{E}[X]$, $\mathrm{E}[X|Y=0]$ and $\mathrm{E}[X|Y=1]$ – the expected value, expected value given S_L is idle and the expected value given S_L is busy of the variable X, respectively.

4.3 System Utilization

We now find the maximum utilization of the system. The *effective-arrival-rate* to S_Q, $\hat{\lambda}(p,\rho)$, consists of two streams of arrivals: customers who join S_Q without sensing (whose

proportion is $1-p$) and customers who sense S_L and find it busy (whose proportion is $p \cdot \Pr(Y=1)$). Therefore,

$$\hat{\lambda}(p,\rho) = (1-p)\Lambda + p \cdot \Pr(Y=1)\Lambda \,.$$

Using (4.4) and dividing by μ we get the *effective-utilization*, $\hat{\rho}(p,\rho)$:

$$\hat{\rho}(p,\rho) := \frac{1}{\mu}\hat{\lambda}(p,\rho) = \rho - p\rho + \frac{p\rho}{1+p\rho}p\rho = \rho - \frac{1}{1+p\rho}p\rho \,, \tag{4.7}$$

which is monotone decreasing in p for every $\rho \in (0,\infty)$. Note, by (4.5), that $\hat{\rho}(p,\rho)$ is equal to $1-(P_{0,0}+P_{0,1})$, so it represents the *busy fraction* of subsystem S_Q.

The stochastic process representing the stream of arrivals to the shared queue is not Poisson when $p>0$. This is because the inter-arrival times to the queue are not i.i.d., and the arrival is higher when S_L is busy than when it is idle. This is a special case of queues with heterogeneous inter-arrival times presented in Yechiali and Naor (1971). For the system to remain stable, as implied by Yechiali and Naor (1971), we assume $\hat{\rho}(p,\rho) < 1$.

Recall the *golden ratio*, $\varphi = \frac{1}{2}\cdot(1+\sqrt{5}) \approx 1.618$. Then:

Proposition 4.1. *The system is stable if and only if* $\rho \in [0,\varphi)$ *and* $p \in (\underline{p},1]$, *where*

$$\underline{p} := \frac{\rho-1}{\rho(2-\rho)} \,.$$

The proof of Proposition 4.1 is in Appendix C.1.

Note that for $\rho<1$, $\underline{p}$, and clearly, the system then is stable for all $p\in[0,1]$. Throughout the paper we assume that $\rho \in (0,\varphi)$ (when $\rho=0$ the solution is trivial). Fig.4.1 shows the domain of possible pairs (ρ,p) such that the system is stable, which is bounded by the curve $\underline{p}$.

4.4 Equilibrium Strategy

In this section we discuss the (Nash) equilibrium strategies in the system. Let $c_w>0$ be the cost per unit time of waiting in the shared queue, and $c_s>0$ be the cost of sensing (incurred by each customer who chooses to sense, regardless of the result of this action). Since all customers are identical, and each one arriving at the system eventually receives service, neither the reward from service nor the time spent in service is relevant. We assume the system is not overloaded (i.e., $\rho \in (0,\varphi)$) and all customers act to reduce their expected cost to minimum.

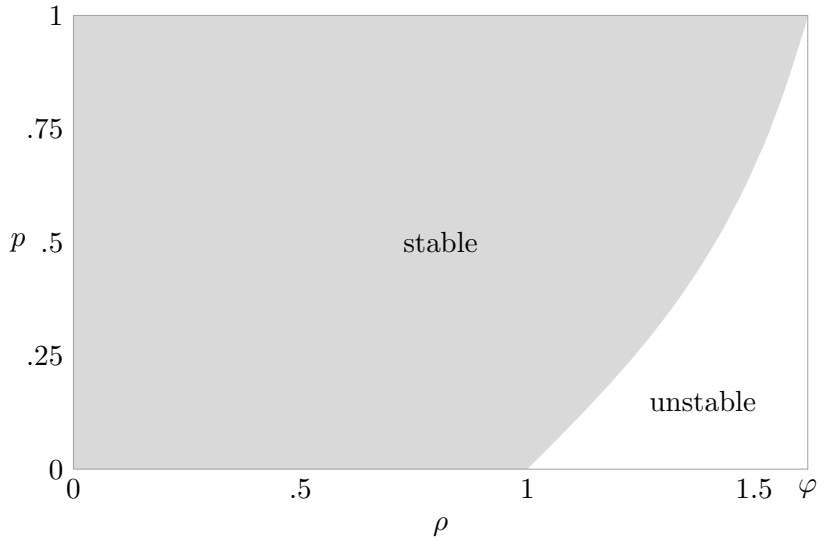


FIGURE 4.1: A colormap showing the p-ρ-plane divided into areas for which the system is stable and unstable. The shaded area in the graph shows values of p and ρ for which the system is stable, which are bounded by the curve $\underline{p}$.

Since the arrival process to the system as a whole is Poisson, the expected number of customers in S_Q upon a customer's arrival is $\mathrm{E}[X]$. The residual service duration of the customer in service in S_Q, and the service duration of all of the other customers awaiting in S_Q are all independent of X and exponentially distributed with mean $\frac{1}{\mu}$. Thus, the expected waiting time of an arbitrary customer in S_Q is $\frac{1}{\mu} \cdot \mathrm{E}[X]$. For a similar reason, given that upon an arrival S_L is busy, the expected waiting time of that customer is $\frac{1}{\mu} \cdot \mathrm{E}[X \mid Y = 1]$

For all $p \in (\underline{p}, 1]$, denote by $C_S(p)$ and $C_N(p)$ the expected cost of an individual who chooses *sense* and *not sense*, respectively, given that the others' sensing probability is p. We have

$$\begin{cases} C_N(p) := \frac{c_w}{\mu} \mathrm{E}\left[X\right] , \\ C_S(p) := c_s + \Pr(Y = 1) \cdot \frac{c_w}{\mu} \mathrm{E}\left[X \mid Y = 1\right] . \end{cases} \tag{4.8}$$

Note that when $p = 0$ then the event $\{Y = 1\}$ is the null set, thus, for $\rho < 1$ we define $C_S(0) = c_s$. Throughout the paper we use the terms $C_N(p)$ and $C_S(p)$ assuming that $p > \underline{p}$. For the avoidance of doubt, we stress that by referring to $p \in [0, 1]$ we mean $p \in [0, 1]$ if $\rho < 1$, and $p \in (\underline{p}, 1]$ when $\rho \in [1, \varphi)$.

We shall now show that there exists a unique equilibrium strategy. To this end, we first point out some properties of $\mathrm{E}\left[X \mid Y = 0\right]$, of which we will make use in the proof:

Proposition 4.2. *The function* $\mathrm{E}\left[X \mid Y = 0\right]$ *is continuous and monotone non-increasing in* p.

The proof of Proposition 4.2 is in Appendix C.2.

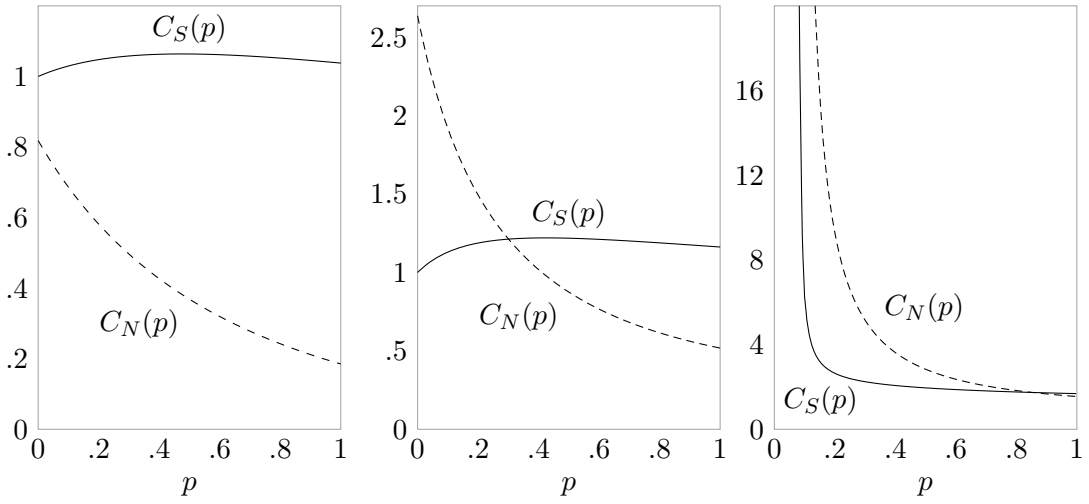
In the proof of Proposition 4.2 it is also shown that $\mathrm{E}[X]$ is monotone non-increasing. The fact that this functions are non-decreasing follows intuition, as the more customers sense the loss system, the smaller the arrival rate to the shared queue, both when $Y = 0$ and at a whole.

Denote $\gamma := \frac{c_w}{(\mu c_s)}$. The value of γ can be interpreted as the normalized cost of waiting a single service period, $\frac{c_w}{\mu}$, paid in currency with rate c_s. Similarly, $\frac{1}{\gamma}$ is the fixed normalized sensing cost.

Proposition 4.3. *For every $\rho \in (0, \varphi)$, and for every $\gamma > 0$, a unique equilibrium strategy $p_e \in [0, 1]$ exists.*

The proof of Proposition 4.3 is in Appendix C.3.

This result is not as intuitive as it seems, since the function $C_S(p)$ is not necessarily monotone (as shown in Fig. 4.2a and 4.2b). This fact can be explained as follows: an increase in the proportion of customers that sense results in decreases probability of attaining service in S_L on one hand. It also decreases the expected queue length, and shortens the waiting time of the customers who failed to attain S_L on the other hand. Nevertheless, the equilibrium probability p_e is unique.

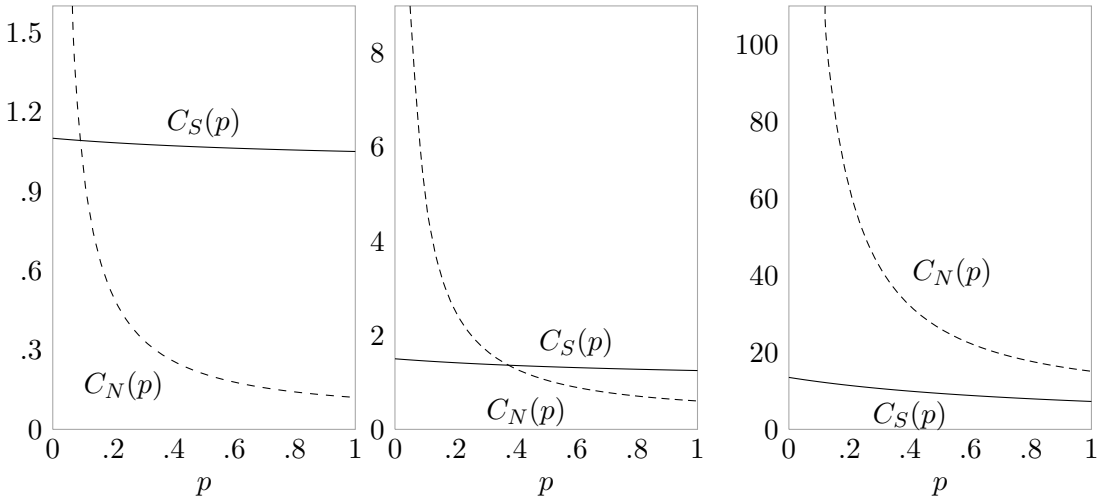


(A) $\gamma = 1$; $\rho = 0.4$; (B) $\gamma = 1$; $\rho = 0.75$; (C) $\gamma = 1$; $\rho = 1.1$;

FIGURE 4.2: The expected costs of sensing (C_S) and not sensing (C_N), normalized in c_s, as functions of p for $\gamma = 1$ and various values of ρ.

Proposition 4.4. *The pure strategy $p = 0$ is an equilibrium strategy (in other words $p_e = 0$) if and only if:*

$$\rho \leq \frac{1}{1+\gamma} .$$



(A) $\gamma = 0.1$; $\rho = 1$; (B) $\gamma = 0.5$; $\rho = 1$; (C) $\gamma = 12.5$; $\rho = 1$;

FIGURE 4.3: The expected costs of sensing (C_S) and not sensing (C_N), normalized in c_s, as functions of p for $\rho = 1$ and various values of γ.

The proof of Proposition 4.4 is in Appendix C.4.

Proposition (4.4) provides a necessary and sufficient condition for $p_e = 0$. We would like to derive a similar condition determining when $p_e = 1$. Define $g(z)$ as

$$g(z) := (1-p)\rho^2 z^3 - (\rho^2 + (3-2p)\rho)z^2 + (2\rho+2)z - 1\,. \tag{4.9}$$

In their work, Yechiali and Naor (1971) have shown that the polynomial $g(z)$ possesses a unique root, denoted z_0, in the interval $z \in (0,1)$, and that:

$$P_{0,0} = \frac{(1-\hat{\rho})z_0}{(1-z_0)(1-\rho z_0)}\,, \tag{4.10}$$

where $\hat{\rho}$ stands for $\hat{\rho}(p,\rho)$, as defined in (4.7).

Define the partial generating function of the system for $Y(t) = 0$ as:

$$G_{Y=0}(z) := \sum_{m=0}^{\infty} z^m P_{0,m}\,, \quad |z| \le 1\,. \tag{4.11}$$

It has been proven by Yechiali and Naor (1971) that

$$G_{Y=0}(z) = \frac{(1-\hat{\rho})z + P_{0,0}(1-z)(\rho z - 1)}{g(z)}\,.$$

Note that

$$\mathrm{E}[X \mid Y=0] = \frac{1}{\Pr(Y(t)=0)} \cdot \frac{d}{dz} G_{Y=0}(z)\Big|_{z=1}\,,$$

hence, denoting $g'(1) := (\frac{d}{dz})g(z)|_{z=1}$,

$$\mathrm{E}[X \mid Y=0] = \frac{(1-\hat{\rho}+(1-\rho)P_{0,0})\cdot g(1) - (1-\hat{\rho})\cdot g'(1)}{\Pr(Y(t)=0)\cdot (g(1))^2}\,. \tag{4.12}$$

Proposition 4.5. *The pure strategy $p=1$ is an equilibrium strategy (in other words $p_e=1$) if and only if:*

$$\gamma \geq \frac{\theta^2+\theta^3-\theta^4}{1-\theta-\theta^2+\theta^3}$$

where $\theta := \sqrt{1+\rho}$.

The proof of Proposition 4.5 is in Appendix C.5.

Proposition 4.6. *The equilibrium strategy p_e is monotone non-decreasing both as a function of ρ and as a function of γ.*

The proof of Proposition 4.6 is in Appendix C.6.

As a matter of fact, the technique of proving Proposition 4.5 is useful in formulating necessary and sufficient conditions not only for $p_e=1$ but also for every value $p_e \in (0,1)$. Put simply, for every set of parameters $\rho \in (0,\varphi)$ and $\gamma > 0$ one can evaluate p_e by substituting (4.12) in (C.13) and solving the equation for p (the solution depends on the root of the polynomial $g(z)$). In particular, when $\rho = 1$, equation (4.12) yields $\mathrm{E}[X \mid Y=0] = \frac{1}{p}$, and applying the conditions for equilibrium it emerges that $p_e = \min\left\{\frac{(\sqrt{1+4\gamma}-1)}{2}, 1\right\}$. This result is reflected in Fig.4.3. Fig.4.4 depicts the value of p_e calculated for various values of ρ and γ.

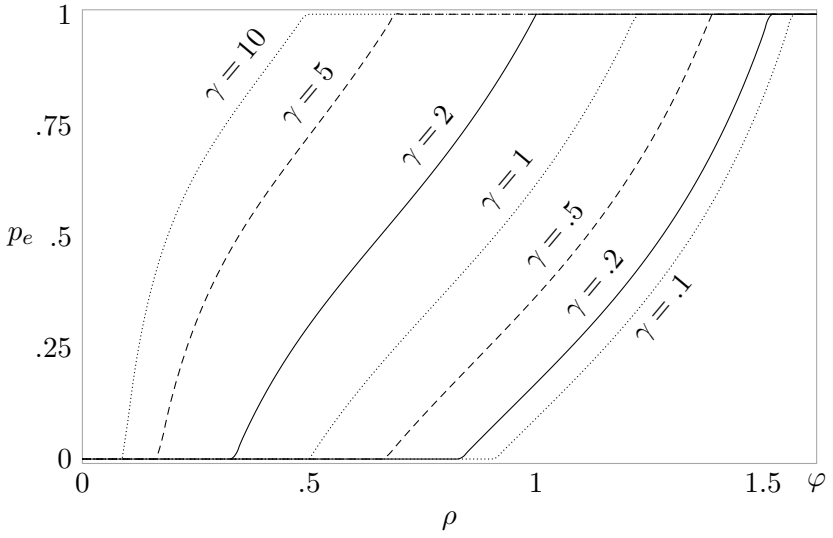


FIGURE 4.4: The equilibrium strategy p_e as a function of ρ for a various values of γ

4.5 Social Optimization

We now turn our attention to social optimization. The social objective function (normalized by $\frac{c_w}{\mu}$), $C(p)$, is defined as

$$C(p) := \frac{\mu}{c_w}\left((1-p)C_N(p) + pC_S(p)\right) , \tag{4.13}$$

and represents the normalized expected cost of a customer when the probability of sensing is p. Combining (4.13) and (4.8) we obtain

$$C(p) = \frac{p}{\gamma} + (1-p)\cdot \mathrm{E}[X] + p\cdot \Pr(Y=1)\cdot \mathrm{E}[X \mid Y=1] . \tag{4.14}$$

Define $F(p)$ as follows:

$$F(p) := (1-p)\cdot \mathrm{E}[X] + p\cdot \Pr(Y=1)\cdot \mathrm{E}[X \mid Y=1] = C(p) - \frac{p}{\gamma} , \tag{4.15}$$

The function $F(p)$ can be interpreted as the expected cost of a customer in a system of which $c_s = 0$ and $\frac{c_w}{\mu} = 1$ (that is to say sensing is free of charge and the expense of waiting for a single service period is 1 on average).

Proposition 4.7. *The function $F(p)$ is convex in p.*

The proof of Proposition 4.7 is in Appendix C.7.

Denote p^* the socially optimal strategy. Accordingly,

$$p^* := \arg\min_{p\in[0,1]} C(p) = \arg\min_{p\in[0,1]} \left\{ \frac{p}{\gamma} + F(p) \right\} . \tag{4.16}$$

From Proposition 4.7 we immediately have that $C(p)$ itself is a convex function, thus, p^* is well defined.

Proposition 4.8.

(a) $p^* = 0$ *if and only if* $\rho \le 1 - \sqrt{\frac{\gamma}{(1+\gamma)}}$.

(b) $p^* = 0 \;\Rightarrow\; p_e = 0$.

(c) *For every* $\gamma > 0$, $\rho \in [1 - \sqrt{\frac{\gamma}{(1+\gamma)}}, \frac{1}{(1+\gamma)}] \;\Rightarrow\; p_e = 0 < p^*$

The proof of Proposition 4.8 is in Appendix C.8.

Proposition 4.8 shows that it is possible that $p_e < p^*$. In fact, there exist values of γ such that for all $\rho \in (0, \varphi)$, $p_e \le p^*$ (see Fig.4.6a) and there exist values of ρ such that

for all $\gamma \geq 0$, $p_e \leq p^*$ (see Fig.4.7b). We shall show next that there exist values of ρ and γ such that $p^* < p_e$. In such cases we say that the system is in a situation of *over-sensing*. Although the value of p^* is computed in a numerical optimization technique, the existence of $p^* < p_e$ can be proved analytically. We begin by formulating a sufficient condition for $p^* < 1$:

Recall $F(p)$ as defined in (4.15).

Proposition 4.9. *If there exists a value $h \in (0,1)$ such that*

$$\gamma \leq \frac{h}{F(1-h) - F(1)}\,, \tag{4.17}$$

then $p^* < 1$.

The proof of Proposition 4.9 is in Appendix C.9.

For a given fixed ρ and h, define:

$$\gamma_1(\rho) := \frac{\theta^2 + \theta^3 - \theta^4}{1 - \theta - \theta^2 + \theta^3}; \quad \gamma_2(\rho) := \frac{h}{F(1-h) - F(1)}\,,$$

where $\theta = \sqrt{1+\rho}$. The relation between these two quantities, $\gamma_1(\rho)$ and $\gamma_2(\rho)$ indicates whether there exists a value of γ such that $p^* < p_e$. Proposition 4.5 states that if $\gamma_1(\rho) \leq \gamma$ then $p_e = 1$. In addition, by Proposition 4.9, if $\gamma < \gamma_2(\rho)$ then $p^* < 1$. We deduce therefore that if $\gamma_1(\rho) < \gamma_2(\rho)$ then for all $\gamma \in [\gamma_1(\rho), \gamma_2(\rho))$, $p^* < p_e = 1$. This remarkable phenomenon can be spotted in Fig.4.6b and Fig.4.7a. In particular, taking $\rho = 0.6$ and $h = 0.01$ satisfies $\gamma_1(\rho) < \gamma_2(\rho)$ as demonstrated in Fig.4.7a. Combining the aforementioned result with Proposition 4.8, we conclude that in some rare cases, $p_e = p^* \notin \{0,1\}$, namely p_e is an efficient non-trivial mixed equilibrium strategy. Such cases are spotted in Fig.4.8, which shows the two-dimensional ρ-γ-plane painted in different colors with respect to each of the three cases: $p^* > p_e$, $p^* = p_e$ and $p^* < p_e$.

4.5.1 Price of Anarchy

The existence of ρ and h such that $\gamma_1(\rho) < \gamma_2(\rho)$ implies that rational customers, under specific conditions, may over-sense S_L. This situation, while possible, may not be a matter of significant importance in practice, as the expected cost of customers in equilibrium is generally only slightly worse than that in social optimum under the circumstances of over-sensing. We reach this observation based upon numerical analysis of the *price of anarchy* (PoA), which is defined as follows:

$$\mathrm{PoA}(\rho,\gamma) := \frac{C(p_e)}{C(p^*)}\,.$$

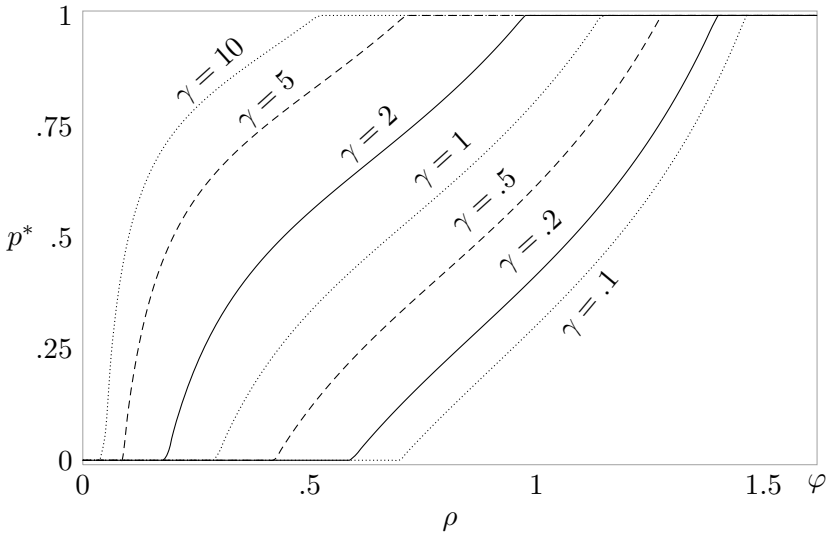
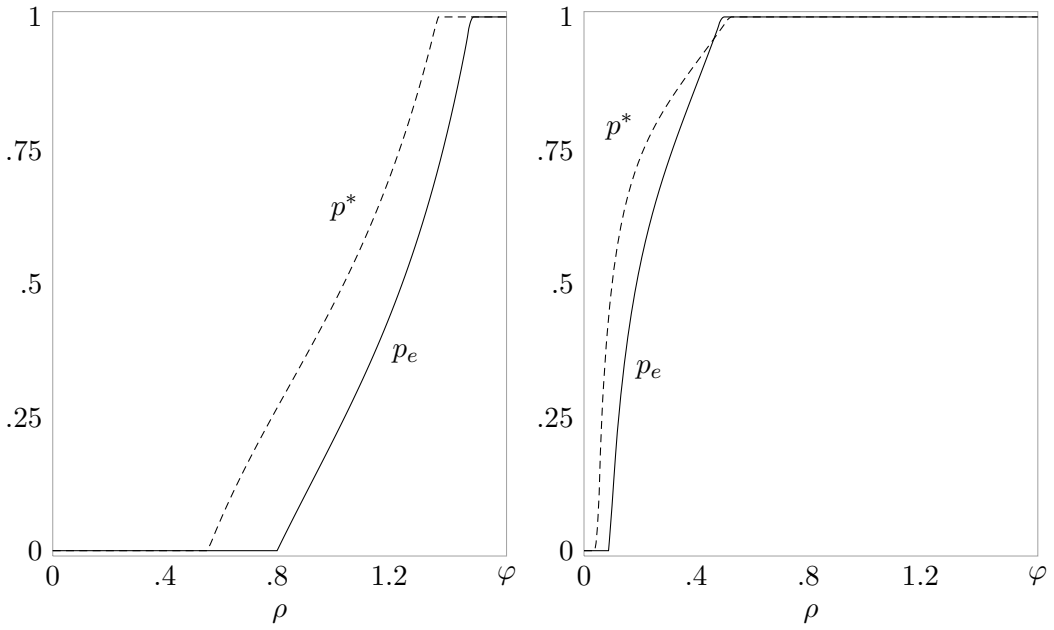


FIGURE 4.5: The socially optimal strategy p^* as a function of ρ for a various values of γ

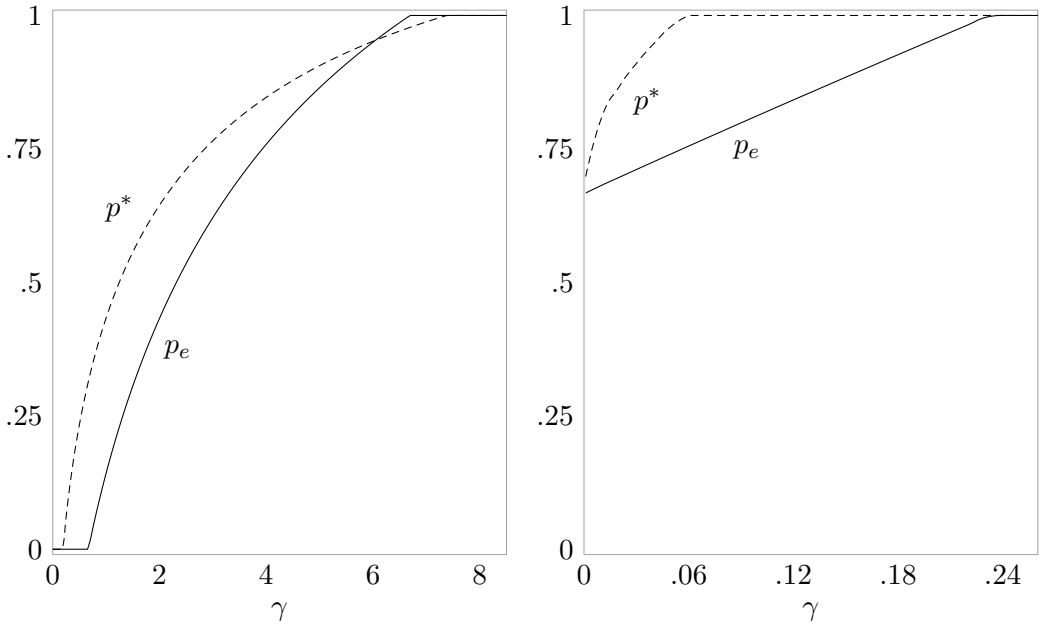


(A) $\gamma = 0.25$; (B) $\gamma = 10$;

FIGURE 4.6: The socially optimal strategy (p^*) and the equilibrium strategy (p_e) as a function of ρ for various values of γ such that $\gamma \leq 1$.

Empirical examination shows that the behavior of PoA as a function of ρ conforms with the results in Proposition 4.8 and Proposition 4.4 (see Fig.4.9). For a fixed value of γ:

- If $\rho \leq 1 - \sqrt{\frac{\gamma}{(1+\gamma)}}$ then $p_e = p^* = 0$ and PoA $= 1$.
- PoA (as a function of ρ) attains its maximum (non-smooth) point at $\rho = \frac{1}{(1+\gamma)}$.



(A) $\rho = 0.6$; (B) $\rho = 1.5$;

FIGURE 4.7: The socially optimal strategy (p^*) and the equilibrium strategy (p_e) as a function of γ for various values of ρ such that $\rho < 1$.

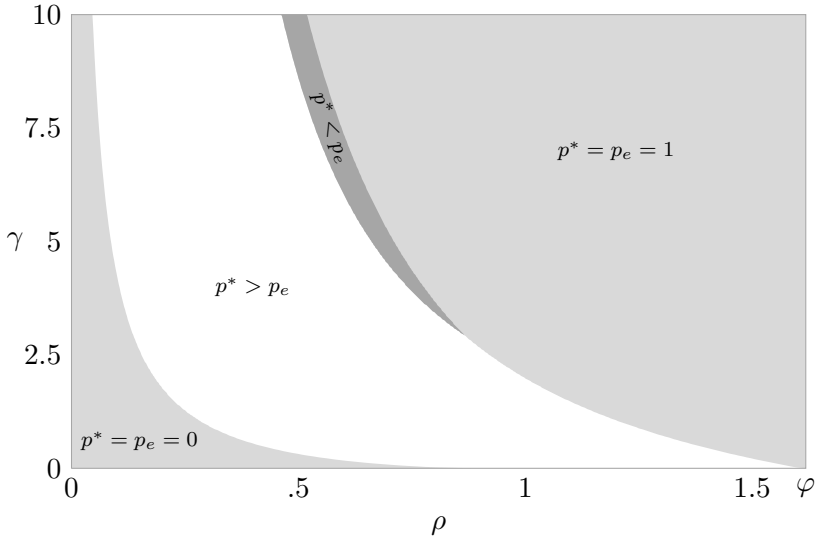


FIGURE 4.8: A color map of the two-dimensional ρ-γ-plane showing the relation between p^* and p_e. For each pair (ρ, γ), white indicates that $p^* > p_e$, gray indicates that $p^* = p_e$ (in particular where they are both equal to 0 or 1) and black indicates that $p^* < p_e$.

where the first and the second properties follow from Proposition 4.8 and Proposition 4.4, respectively.

For a fixed given γ, we define ρ_1 such that $\gamma_1(\rho_1) = \gamma$. If $\gamma_1(\rho_1) < \gamma_2(\rho_1)$, then taking $\rho = \rho_1$ and γ to be the system parameters implies that $p^* < p_e = 1$. For this fixed value of γ, we notice that PoA is not a unimodal function of ρ, and attains a local

maximum (non-smooth) point at ρ_1 in addition to its maximum point at $\frac{1}{(1+\gamma)}$. This can be observed in Fig.4.10, where $\gamma = 10$. Note that in the case where $\gamma = 10$ (see Fig.4.6b) then $\rho_1 \approx 0.49$, it holds that

$$\gamma_1(\rho_1) \approx 10.07 < \lim_{h \to 0} \gamma_2(\rho_1) \approx 11.23$$

and $\text{PoA}(0.49, 10) = 1.0005$, which means that $C(p_e)$ is marginally greater than $C(p^*)$.

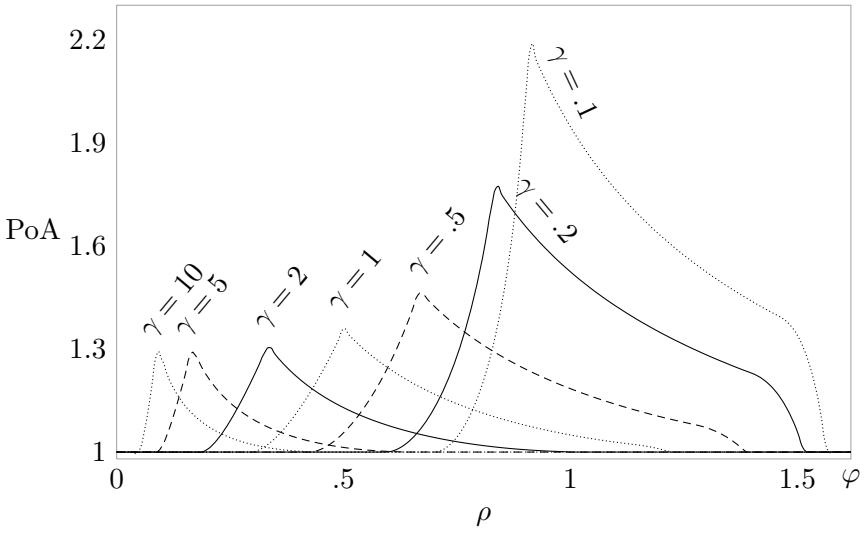


FIGURE 4.9: The Price of Anarchy (PoA) as a function of ρ for a various values of γ

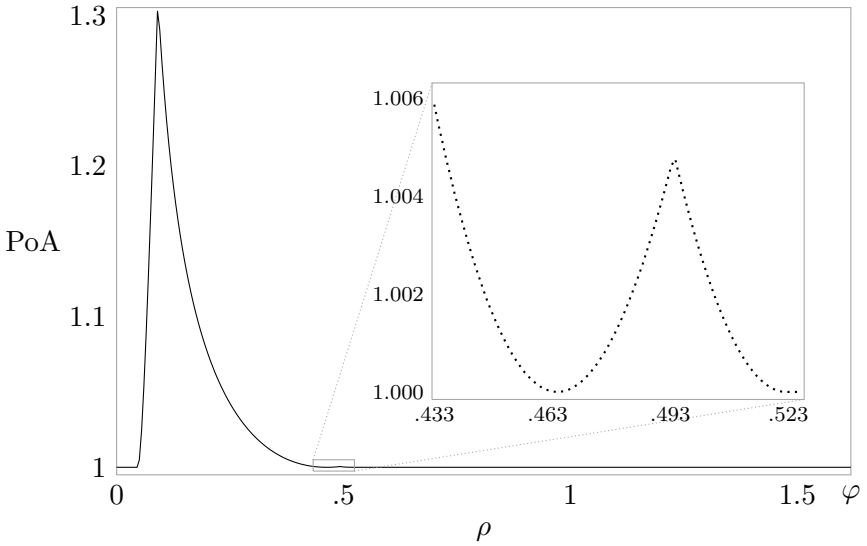


FIGURE 4.10: The Price of Anarchy (PoA) as a function of ρ for $\gamma = 10$. This example corresponds with Fig.4.6b, where $p^* = p_e$ (i.e., PoA = 1) at $\rho \approx .465$, for which the equilibrium strategy point is efficient and non-trivial.

4.6 Concluding Remarks

This work deals with customer behavior in a queueing model that involves choosing between a loss system and a system of an infinite buffer. From the queueing analysis aspect, the main challenge is handling the dependency between the arrival rate to the two subsystems, S_Q and S_L and the customers' actions. In this specific model, customers' actions induce an arrival process to S_Q which in general is not Poisson.

We manage to show that the equilibrium strategy in this multi-player game is unique, and can be either greater, smaller or equal to the socially optimal strategy. This results continue to hold for even when the service rates of S_Q and S_L are different, because the proof depends mainly on the fact that the service duration in both servers is memoryless and on the fact that the loss-probability of S_L is monotone increasing in p. Only slight changes are needed in order to apply the same solution for a network with different service rates, and the differences are primarily technical.

By proving properties (monotonicity and convexity) of the conditionally expected queue length and expected cost (which are functions of the sensing probability) we show the uniqueness of the equilibrium strategy, p_e, and compare it with the socially optimal one, p^*. However, the basic tools of calculus are of no use as these functions are given in implicit form. In this chapter, specifically in the proof of Proposition 4.2 and Proposition 4.7, we present a construction of an appropriate coupling argument that appear to be essential for the analysis of the model. We have further shown that there are cases where $p* \geq p_e$ and cases where $p* < p_e$, depending on the system parameters. If indeed $p^* < p_e$ then in order to reach optimal sensing rate customers should be incentivized to join the shared queue without sensing. We find this an unusual result, because in many models such as in Roet-Green and Hassin (2014), in the mean field model in Xu and Hajek (2013), and in the fundamental model of Edelson and Hildebrand (1975), the proportion of customers joining the queue in equilibrium is never less than the socially optimal proportion. Moreover, the act of purchasing information in Roet-Green and Hassin (2014) and in the mean field model in Xu and Hajek (2013) induces positive externalities on society that a customer does not take into consideration when aiming for individual optimization. As for the model presented here, this argument in general does not hold, and the explanation is as follows: Say two successive customers arrive at the system one after another. Whereas in most models it is beneficial for the second customer that the first decides not to join the queue, in this model it is not clear whether the second one prefers that the first one senses or joins.

One more way for addressing the over-sensing phenomenon is by observing the PoA, which can be expressed as a function of the system utilization, ρ. When ρ is small

relative to γ we have $p_e = p^* = 0$, and when ρ is relatively large we have $p_e = p^* = 1$. By definition, if $p_e = p^*$ then PoA $= 1$, otherwise PoA > 1. When increasing ρ, at some point the system may switch from $p_e < p^*$(under-sensing) to $p_e > p^*$ (over-sensing). At this point satisfying $p_e = p^* \in (0, 1)$, the function PoA attains its local minimum. As a consequence, PoA is not a unimodal function of the system utilization.

Further investigation and discussion is needed in order to apply the results in this work on models where there can be more than one server to sense. Extending the number of servers so that S_L becomes an M/M/s/s queue or adding a finite buffer of waiting slots in front S_L so it becomes an M/M/1/K, make the solution of Yechiali and Naor (1971) for the stationary probabilities no longer suitable. Also, if there is a buffer in front of S_L it is necessary to specify what information is revealed to a sensing customer, and the set of actions for customers should be adjusted accordingly. For example, if sensing reveals the exact number of customers awaiting in S_L, then it is not clear that a sensing customer when observing the queue length in S_L would wish to join immediately. To the extent that S_L is a multi-server loss system, then we need to model the sensing methodology: how many servers are sensed, how they are sampled among the collection of servers, etc. If we consider a case where each sensing customer is blocked if and only if the loss system servers are all busy, then it can be shown in similar methods to those presented here, that the equilibrium strategy is unique.

Chapter 5

Tipping in Service Systems: The Role of a Social Norm

5.1 Background and Motivation

On an average day, 10% of the US population leaves a tip in the restaurant they visit (Lynn, 2006). Customers also leave tips for barbers, chambermaids, delivery people, and many more (Star, 1988). In 2006, the magnitude of tips in the restaurant industry in the US alone was estimated at about $44 billion (Azar and Tobol, 2008). Paying a tip is therefore a significant part of any customers' budget to acquire a wide range of services. At the same time, these tips are a crucial source of income for more than two million employees in the US food and beverage sector, paid at (or below) the minimum wage for tipped employees[1] (Jones, 2016).

Tipping is a widespread custom in many societies, but, not in all. Originated in Europe in the sixteenth century and exported to the US in the late 1800s, tipping nowadays is much more prevalent in the US than in Europe (Azar, 2004a). Yet, tipping has been controversial; in the early 1900s, seven US states, have voted and repealed anti-tipping laws, Azar (2005a). Recently, there is a tipping debate raging the US. Bringing up considerations such as fairness, discrimination and theft[2], some employers in the

[1]The minimum wage is $2.16 per hour for employees earning at least $30 in tips per month. This wage plus tips must exceed the maximum of the federal minimum wage of $7.25 per hour (since 2009) and the state minimum wage, see United States Department of Labor, Wage and Hour Division, Fact Sheet #15.

[2]Under current Department of Labor regulations in the US, the tip is owned by the employee and cannot be appropriated by employers, Shierholz et al. (2017). Therefore, tips cannot be shared either with employees who do not customarily and regularly receive tips, such as dishwashers, cooks, chefs, and janitors Wood (2017). 'Tip pooling,' i.e. splitting tips between employees who customarily and regularly receive tips such as waiters/waitresses, bellhops, counter personnel (who serve customers), bussers, and service bartenders is permitted under the Fair Labor Standards Act.

restaurant industry switched from tipping to no tipping, and some switched back to tipping (Yagoda, 2018).

We do not have space here to provide a full overview of the literature on tipping, see Lynn (2006), Azar (2005a) for excellent overviews. In short, economists have focused on the benefits of tipping for employers via (a) reducing employee monitoring costs, Bodvarsson and Gibson (1997), Lynn and McCall (2000), Pencavel (2015), Jacob and Page (1980), (b) facilitating tax evasion, (c) facilitating screening of employees, Schotter (2000) (d) attracting customers, Schwartz (1997), Lynn and Wang (2013) and (e) addressing service delivery failures, Sisk and Gallick (1985).

The benefits of tipping for employers can only be accrued in case that customers are actually motivated to tip. Tips are usually paid after receiving service. For repeat customers, tipping is rational, as these customers may see a return to their tip for services that will be rendered in the future (Ben-Zion and Karni (1977)). However, tipping is not rational in the case that the customers are one-time customers. These customers will never enjoy any return to their tip in the future as they will never re-visit the facility. Hence, assuming that tipping is merely driven by future service considerations, *rational one-time customers should never tip*. Nevertheless, in practice, many such one-time customers do leave a tip. In order to explain one-time customers' tipping behavior, various *behavioral* motivations have been brought forward in the literature; conforming to a social norm is considered as one of the most significant motivations for leaving a tip, Azar (2004a).

The theoretical literature on tipping with a social norm has been pioneered by Azar; see Azar (2007) for a review. Azar (2004b) develops a model in which he endogenizes the social norm keeping the service value fixed; customers incur dis-utility when they tip a different amount than the social norm prescribes. In addition, customers enjoy positive utility from the tipping process as well as from tipping more generously than the norm. Azar finds that without the latter utility, the social norm erodes over time. Azar (2008) endogenizes the service quality in a game between a service provider (determining effort for service) and a customer (determining the tip), in which the social norm is exogenous. Azar finds that when the social norm is very sensitive to quality, tips increase the social welfare, but, that the service quality with tipping may be below the socially optimal value.

A stream of papers studies the empirical relationships between tipping and service quality and tipping and patronage frequency. Repeat customers tip more than one-time or infrequent customers (Lynn and McCall (2000), Conlin et al. (2003)). It is not clear, however, whether the repeat customers receive better service as the positive impact of service quality on tips is somewhat weak. May (1978) found that tips are independent

of service quality. On the contrary, Lynn and Grassman (1990) found a significant positive correlation between tips and patronage frequency, but not between tipping and the interaction of patronage frequency with service ratings, see also Bodvarsson and Gibson (1994). These mixed findings call for a better understanding of the effect of patronage frequency on tips, Azar (2007). The relationship between service rating and tip amount depends critically on the customer's perception of the employee's control over the service, Azar (2004a). In some cases the latter is clear. For food-delivery services, for example, the dimension of service quality is mainly the service speed, which can be attributed more easily to the drivers. Seligman et al. (1986) find that indeed, food delivery persons receive larger tips for faster delivery.

Interestingly, the origins of the word 'tip' dates back to the sixteenth century in England, where brass urns were placed in coffee houses and local pubs with the inscription: 'To Insure Promptitude' or 'To Insure Promptness', Azar (2005a). The English word *waiter* finds its origins in the late 15th century and refers originally to an "attendant at a meal, servant who waits at tables"[3], indicating that the server *waits* for a service request by the guest, who are served promptly. Nowadays, in many service settings human labor is still a constraining resource and service speed is an important (and objective) measure of service quality. Yet, none of the above theories considers service speed as key measure of service quality. As customers typically dislike waiting in practice, we focus in this chapter on tipping for fast service.

A sizable literature in Operations Research studies the impact of service speed on customer behavior in resource-constrained service settings (see Hassin and Haviv (2003), Hassin (2016) for excellent overviews of the literature). In essence, in this literature, an important focus is on the customer-to-customer interaction as they all compete to gain speedy access to a limited resource (the server). Wait-sensitive customers can influence the speed at which they receive service by obtaining 'priority'. A scheme that is the closest related to tipping, is referred to as *bribery*, Lui (1985)[4]. In such scheme, each customer chooses the amount she wishes to pay for priority (the tip) and is then placed in the queue ahead of those who paid less. Therefore, customers paying more have to wait less for service. In a priority queuing setting Glazer and Hassin (1986) specify the equilibrium distribution of bribes (or tips). Specifically, when customers are homogeneous, the server's profit achieves the maximum social welfare and all customers obtain zero utility. Another mechanism achieving the same, maximum social welfare for homogeneous customers, but, with a First Come First Served (FCFS) service discipline instead of a priority-based discipline, is obtained by charging an (optimized) 'admission fee' to all customers who want to join, as in Edelson and Hildebrand (1975). Again, the

[3]See the Online Etymology Dictionary entry for waiter.
[4]See Azar (2005a) for a discussion about 'bribery tipping'.

admission fee extracts all customer surplus. In queuing games literature, customers are assumed to be rational, and, make their decision on priority purchasing *before* receiving service. As discussed above, one-time customers tip *after* having received service and such observation can be explained through the existence of a behavioral component; a social norm. There is an emerging literature, in which behavioral elements are incorporated in queuing games. For example, Huang and Chen (2014) study pricing of services when expectations about waiting times are formed based on past experiences and anecdotal reasoning. Yang et al. (2018) study pricing of services when customers are loss-averse. To the best of our knowledge, no priority queuing game has been analyzed to study the formation of a social norm.

Two questions in the literature regarding tipping are important: Why do one-time customers tip, and, does tipping improve social welfare? We address these questions in this chapter. In line with Bodvarsson and Gibson (1997), customers *first* interact with each other for access to a limited resource in a 'service consumption' stage and *second*, customers interact with each other in a 'social consumption' stage, like Azar (2008), in which they compare how much they tipped. In our model, the social norm is endogenously formed. To keep focus, as in Azar (2004b), we do *not* address the employee's 'effort' to provide 'good service' (see e.g. Jacob and Page (1980), Azar (2008), for such studies with endogenous effort). In our model, the service rate determines the effort in our queuing model and is thus fixed. Good (or speedy) service for one customer who received high priority comes at the expense of slow service for another customer who received low priority. This notion of priority is consistent with Lynn et al. (2012), who suggest that servers should exert more efforts in serving those dining parties they expect to pay larger tips. As in the literature, we study a mixture of repeat and one-time (non-repeat) customers. To make insights sharp, we do not consider any other source of heterogeneity in the customer population. Customers in our model are rational in the sense that they decide whether to purchase the service or not and, if they do, how much to tip, maximizing their expected utility, taking both service and social consumption into account.

We describe our model more formally next.

5.2 Model Description

In this section, we set up first our model and then define equilibrium.

5.2.1 Set-up

We introduce a tipping game which is divided into two stages. In the first stage, the *consumption* stage, delay-sensitive customers arrive according to a Poisson process at a (stationary) single-server service facility. The service value is U, and service time is exponentially distributed with mean $1/\mu$. Customers (a) decide whether to obtain service (join) or not, (b) tip in case they join. In case customers join, a queueing process is formed. The joining decision is taken without knowledge of the actual queue length (as in Edelson and Hildebrand (1975)). We assume customers incur costs while waiting, both in line and during service. The waiting cost rate of all customers is $c > 0$ per unit-time. To avoid trivialities, we assume that $U > c/\mu$, otherwise the service facility does not generate any value even in the absence of congestion. The utility of not joining is normalized to zero. As we explain below, a higher tip will lead to lower expected waiting time and will be compared with other customers' tip in the second stage.

In the second stage, customers who joined, meet in a *social market* with another, randomly selected customer who obtained service. Both customers reveal their tip amount to each other and incur a dis-utility that is proportional to the squared difference of the tip; $\kappa(t - t')^2$ when the revealed tip amounts were t and t', for some $\kappa \geq 0$. The parameter κ captures the strength of the social norm, whose impact on tipping behavior is our main focus.

Consistent with practice we assume that customers tip *after* the service was provided. However, as in Azar (2008), we introduce the notion that some customers (*r*epeat customers) can influence wait time via their tip, while other customers (the *o*ne-time customers) cannot. The potential arrival rate of repeat (one-time) customers to the service facility is Λ_r (Λ_o). Similar to Afèche et al. (2015), we assume that the *individual demand rate* of a repeat customer is sufficiently small with respect to Λ_r such that it does not change the other customers' utility. Nevertheless, the individual repeat customer visits the facility frequently enough such that the server recognizes her, and the server can predict how much she will tip. In line with Ruffle (1999), another interpretation of repeat customers in our model is that they tip the server up front, such that upon joining the facility, the server knows her tip. When a one-time customer joins, the server forms a belief about her tip. The server then gives preemptive priority to the customer with the highest (believed) tip. Customers with exactly the same expected tips are treated FCFS.

All parameters Λ_r, Λ_o, μ, c, V and κ are common knowledge, and customers are rational in the sense that each one seeks to maximize her own individual utility.

The customer selects an action $a \in \{\text{join, balk}\}$ in the first stage. We allow for randomization over the joining decision and let α_χ be the probability that a customer type χ joins. We also allow for randomization of the tipping decision of the customers that joined. The tip amount of customer type χ, can be described via a cumulative distribution $T_\chi(t)$ for $t \in \mathcal{T}_\chi (\subseteq \mathbb{R}^+)$. We that the servers' belief about the one-time customers' tip is correct. We elaborate the equilibrium conditions we impose on $(\alpha_r, \alpha_o, T_r, T_o)$ next.

5.2.2 Equilibrium

For convenience of notation, instead of using α_χ, we introduce the customers effective arrival rate via $\lambda_\chi = \alpha_\chi \Lambda_\chi \in [0, \Lambda_\chi]$, where $\chi \in \{o, r\}$. We denote the unconditional (both one-time and repeat customers) cumulative tipping distribution by $T(t)$ over $\mathcal{T} = \mathcal{T}_o \cup \mathcal{T}_r$. As the server forms beliefs about the one-time customers tip, the cost for a customer at the consumption stage relies on the tipping distribution conjectured by the server, which we assume is correct. Therefore, we use the same notation $T(t)$ for both the conjectured and the actual tipping distribution. The expected utility of customer type χ is

$$U_\chi(t_\chi, T, \lambda_o, \lambda_r) = U - \underbrace{t_\chi}_{\text{Tip}} - \underbrace{cW_\chi\left(t_\chi, T, \lambda_o, \lambda_r\right)}_{\text{Waiting cost}} - \underbrace{\kappa \int_{\tau \in \mathcal{T}} (\tau - t_\chi)^2 dT(\tau)}_{\text{Social cost}} \tag{5.1}$$

where W_χ is the expected waiting time. For a repeat customer, the expected waiting time, W_r, depends directly on the tip, t_r, as well as on the distribution of tip payments and on the customers' joining rates, λ_o, λ_r. For one-time customers, W_o does not depend on their tip, but, does depend on the server's belief about the tip, the joining rates and the tipping distribution. For both customer types, the social cost does depend on the tipping distribution.

As customers are homogeneous, type-χ customers should expect the same utility of joining. Thus, in equilibrium the expected utility for all tips in the randomization domain, $\mathcal{T}_\chi$, equals to u_χ. For tips outside the randomization domain, the expected utility in equilibrium is lower than u_χ. As customers are rational and the utility of not joining is normalized to zero, when u_χ is strictly positive (negative), χ-type customers do (not) join, otherwise, they randomize. Thus, for type $\chi \in \{o, r\}$ with joining rate

$\lambda_\chi \in [0, \Lambda_\chi]$, the equilibrium conditions are:

$$\begin{cases} \text{rational tipping:} & u_\chi = U_\chi(t, T, \lambda_o, \lambda_r) \text{ for all } t \in \mathcal{T}_\chi \text{ and} \\ & u_\chi \geq U_\chi(t, T, \lambda_o, \lambda_r) \text{ for all } t \notin \mathcal{T}_\chi, \\ \text{rational joining:} & \lambda_\chi = 0 \text{ when } u_\chi < 0,\ \lambda_\chi = \Lambda_\chi \text{ when } u_\chi > 0 \text{ and} \\ & \lambda_\chi \in [0, \Lambda_\chi] \text{ when } u_\chi = 0. \end{cases} \tag{5.2}$$

For each type $\chi \in \{o, r\}$, let the joining rate λ_χ^* with tipping domain $\mathcal{T}_\chi^*$ satisfy the conditions above, and let $T^*(t)$ be the unconditional tip distribution. For the convenience of notation, we drop the * indication and, throughout the paper, unless stated otherwise, we refer only to equilibrium strategies.

5.3 Analysis

5.3.1 Preliminary Results

We start with a proposition about equilibrium joining and tipping behavior (Proposition 5.1) that will determine subsequent analysis.

First of all, even though all repeat customers are homogeneous, their tipping distribution (excluding one-time customers' tips) must be non-degenerate: Recall that customers are served in an FCFS manner when they tip the same amount. Suppose on the contrary that a non-infinitesimal proportion of repeat customers tip the same amount. A single repeat customer who deviates from that amount by tipping an arbitrarily small amount $\varepsilon > 0$ higher, will skip over a non-zero, ε-independent number of customers. Hence, such deviation is attractive because it yields a substantial improvement in waiting time. As a consequence, repeat customers cannot all tip the same amount in equilibrium. Therefore, there must exist a non-empty domain over which repeat customers randomize their tip. This is similar to Glazer and Hassin (1986). For one-time customers, however, the situation is different. As tipping does not impact their priority directly, the one-time customers' tip minimizes the social plus tipping cost, which is identical for all one-time customers. Thus, the tipping density of one-time customers is degenerate.

Next, we note that the tip amount chosen by one-time customers is equal to the lowest tip in the repeat customers tipping domain. Assume on the contrary that the one-time customers' tip is strictly lower than the lowest tip of the repeat customers. Consider a repeat customer who tips the least among all other repeat customers. If she deviates by tipping strictly more than the one-time customers' tip, but, strictly less than all other repeat customer's tip, she will obtain exactly the same waiting time. Yet, she will

be strictly better off, as her tip becomes closer to the one-time customers' tip, which is the minimizer of the tipping plus social norm cost. Therefore, there cannot be any measurable 'gap' between the one-time customers' tip and the repeat customers' tipping domain.

Finally, in the presence of one-time customers, repeat customers will always tip strictly more than one-time customers, because, as explained, a repeat customer who chooses exactly the same amount as the one-time customers, will be strictly better off tipping slightly more. Hence, all repeat customers have the same utility (as they randomize) and are strictly better off in any equilibrium than one-time customers. We obtained:

Proposition 5.1. *In equilibrium:*

(i) The tipping distribution of repeat customers is non-degenerate, continuous, and strictly increasing, and its domain is an interval $\mathcal{T}_r^* = (\underline{t}, \bar{t}]$.

(ii) The tipping distribution of one-time customers is degenerate $\mathcal{T}_o^* = \{\underline{t}\}$ *and is equal to the lowest tip of repeat customers.*

(iii) Repeat customers obtain strictly higher utility in equilibrium than one-time customers.

The proof of Proposition 5.1 is in Appendix D.1.

Proposition 5.1 items (i-ii) imply that in equilibrium, one-time customers tip the lowest among all customers, regardless of their actual waiting time and hence, will receive the lowest priority for service. It also follows that we only need to characterize the tipping distribution for repeat customers, T_r, and its domain, $[\underline{t}, \bar{t}]$, with $\underline{t}$ being the tip paid by one-time customers. Let $\lambda = \lambda_r + \lambda_o$ denote the total arrival rate that joins. The unconditional tipping distribution T is a mixture of the distribution T_r with probability[5] λ_r/λ and (the degenerate distribution at) $\underline{t}$ with probability λ_o/λ, thus, containing an atom at $\underline{t}$ when $\lambda_o > 0$.

A consequence of Proposition 5.1(iii) is that one-time customers join the system in equilibrium if and only if all the potential demand of repeat customers join with strictly positive utility. There are thus three possible outcomes in equilibrium:

(i) A fraction of repeat customers join; $\lambda_r \in (0, \Lambda_r]$ and $\lambda_o = 0$. All customers receive zero expected utility.

[5] As the individual joining rate of a repeat customer is small compared to Λ_r (and λ_r) the probability she is paired with herself during the social consumption stage, is negligible and therefore, the probability of meeting another repeat customer is $\lambda_r/(\lambda_o + \lambda_r)$.

(ii) All repeat customers and a fraction of one-time customers join; $\lambda_r = \Lambda_r$ and $\lambda_o \in (0, \Lambda_o)$. Repeat customers receive strictly positive expected utility, and one-time customers receive zero expected utility.

(iii) All (repeat and one-time) customers join; $\lambda_o = \Lambda_o$ and $\lambda_r = \Lambda_r$. Repeat customers receive strictly positive expected utility and one-time customers receive non-negative expected utility.

To keep focus, we assume that the potential arrival rates satisfy $\Lambda_r \in (0, \mu)$ and $\mu < \Lambda_r + \Lambda_o$. With this assumption, Case (iii) can never be an equilibrium. So, the total arrival rate, λ, uniquely determines the equilibrium arrival rates of each type: $\lambda_r = \min\{\lambda, \Lambda_r\}$ and $\lambda_o = \{\lambda - \Lambda_r\}^+$.

Combining Proposition 5.1 and $\mu < \Lambda_r + \Lambda_o$ altogether implies that in the equilibrium characterized by Equation (5.2), the customer who tips the least amount $\underline{t}$, whether it is a repeat customer ($\lambda < \Lambda_r$) or it is a one-time customer ($\lambda > \Lambda_r$), receives zero expected utility. Therefore, in what comes next, we solve for the equilibrium arrival rate λ and the corresponding tip distribution T, by applying the following process: First, we assume that the joining rate λ is known, and specify the corresponding tipping distribution $T(t;\lambda)$ and its domain $[\underline{t}, \bar{t}]$, such that for this given rate, repeat customers are indifferent between tipping every $t \in (\underline{t}, \bar{t}]$ (but strictly prefer it to tipping $t' \notin (\underline{t}, \bar{t}]$). Then, as a function of λ, we compute the total cost (tip, wait and social cost) of the customer tipping $\underline{t}$, for the induced distribution $T(t;\lambda)$. Finally, we solve for λ satisfying that this total cost equals the service value, U.

5.3.2 Tipping without a social norm

In this subsection, before analyzing the impact of the social norm ($\kappa > 0$) on the tipping strategy, we first characterize the equilibrium in the absence of a social norm ($\kappa = 0$). Proposition 5.2 prescribes the equilibrium tipping distribution and the waiting times, assuming the joining rate, λ is known. Next, we characterize T_r, $\underline{t}$, $\bar{t}$ and the corresponding expected waiting times:

Proposition 5.2. *In the absence of a social norm ($\kappa = 0$): For given $\lambda \in (0, \mu)$:*

(i) A unique equilibrium cumulative distribution function of repeat customers' tips is given by

$$T_r(t;\lambda) = \frac{\mu}{\min\{\lambda,\Lambda_r\}} \frac{1}{\sqrt{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\mu}{c}t}} + 1 - \frac{\mu}{\min\{\lambda,\Lambda_r\}}$$
$$\text{for } t \in (0, \bar{t}(\lambda)], \text{ with } \bar{t}(\lambda) = \frac{c}{\mu}\{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - 1\},$$

and is strictly convex increasing in t.

(ii) The expected waiting time as a function of the tip amount for repeat customers is:

$$W_r(t;\lambda) = \frac{1}{\mu(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{t}{c}$$

and for one-time customers is

$$W_o(\lambda) = \frac{1}{\mu(1-\frac{\lambda}{\mu})(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})}.$$

The proof of Proposition 5.2 is in Appendix D.2.

Consistent with e.g. Seligman et al. (1986), Lynn et al. (2012), from Proposition 5.2(ii), it follows that the waiting time (service quality) decreases (increases) in the tip amount; W_o (for one-time customers tipping $\underline{t}$) is greater than $W_r(t)$ (for repeat customers tipping $t > \underline{t}$) and $\frac{d}{dt}W_r(t) < 0$. In addition, repeat customers obtain faster service than one-time customers (when they join), which is intuitive and consistent with Ben-Zion and Karni (1977). Here, in order to make the repeat customers indifferent over the tipping domain, the expected waiting cost ($cW_r(t)$) must decrease linearly in the tip (such that the expected tipping plus waiting costs become independent of the tip); $\frac{d}{dt}W_r(t) = -1/c$.

Interestingly, the cumulative distribution function of tips for repeat customers is convex. The intuition is the following: From priority queuing (e.g. Kleinrock (1976)), we know that when repeat customers' tipping distribution is T_r, the expected waiting time for a repeat customer tipping t is proportional to $(1-\frac{\lambda}{\mu}+\frac{\lambda}{\mu}T_r(t))^{-2}$, which is a convex decreasing function of $T_r(t)$. Consider two tipping amounts, $t_p < t_q$ with $p = T_r(t_p) < q = T_r(t_q)$. The marginal waiting cost saved by a customer who moves from tipping t_p to tipping $t_{p+\varepsilon}$ (the $(p+\varepsilon)$-quantile) is greater than the cost saved by a one who moves from tipping t_q to tipping $t_{q+\varepsilon}$. Thus, in order for the waiting cost to provide perfect compensation for the tip, it must follow that the difference in tip amount between the t_p and $t_{p+\varepsilon}$ is bigger than the difference between t_q and $t_{q+\varepsilon}$. Hence, the cumulative distribution must be convex, which, in turn, implies that the tip density is increasing. It further implies that t_p is convex increasing as a function of p, and also as a function of the associated waiting cost, $W_r(t_p)$ – we rely on these facts later in §5.3.3.

Finally, notice that the lowest tip among for the repeat customers is zero. Therefore, if one-time customers join, their tip is also zero. This is also intuitive: Assume that the latter was strictly positive, say $0 < \underline{t}$. In that case, a customer who tips $\underline{t}$, by tipping less, will not change her expected waiting time but clearly increase her utility. Hence, such deviation from equilibrium would be profitable. It follows immediately from Propositions 5.1 and 5.2 that:

Corollary 5.3. *In the absence of a social norm ($\kappa = 0$), one-time customers never tip ($\underline{t} = 0$) when they join ($\lambda > \Lambda_r$).*

Not surprisingly, as one-time customers cannot influence service priority via their tip, in the absence of social norm, one-time customers never get any benefit from tipping, therefore they do not tip. This implies that externalities caused by repeat customers competing for access to the same resources do not motivate rational one-time customers to tip. In §5.3.3, we study whether a social norm will make these customers tip (or not).

With Proposition 5.2, we can now characterize the equilibrium demand rate, λ. This, as explained in §5.3.1, is done by equating the total cost of the customer tipping $\underline{t}(= 0)$, to the service value, U. We denote by $C(\lambda)$ the total expected cost for that customer:

$$C(\lambda) = \begin{cases} cW_r(0;\lambda), & \lambda \in (0, \Lambda_r], \\ cW_o(\lambda), & \lambda \in (\Lambda_r, \mu). \end{cases}$$

Now, Equation (5.2) becomes:

$$U = C(\lambda),\ \lambda \in (0, \mu).$$

We next characterize the equilibrium joining behavior:

Proposition 5.4. *In the absence of a social norm ($\kappa = 0$):*

(i) When $\Lambda_r \in (0, \mu - \sqrt{c\mu/U}]$, then all repeat customers and some one-time customers join in equilibrium ($\lambda = \mu - c/(U(1 - \frac{\Lambda_r}{\mu})) > \Lambda_r$). Repeat customers tip according to $T_r(t;\Lambda_r)$, over $(0, \bar{t}(\Lambda_r)]$. The one-time customers tip 0.

(ii) When $\Lambda_r \in (\mu - \sqrt{c\mu/U}, \mu)$, then not all repeat customers join in equilibrium, $\lambda = \mu - \sqrt{c\mu/U} < \Lambda_r$. Repeat customers who join tip according to $T_r(t; \mu - \sqrt{c\mu/U})$, over $[0, U - c/\mu]$. The one-time customers do not join.

The proof of Proposition 5.4 is in Appendix D.3.

The term $\mu - \sqrt{c\mu/U}$ is also the socially optimal arrival rate in any (unobservable queuing) system with potential arrival rate exceeding the service rate, as under our assumption $\mu < \Lambda_r + \Lambda_o$, see Edelson and Hildebrand (1975), Glazer and Hassin (1986). Hence, according to Proposition 5.4, when the potential market for repeat customers is lower than the socially optimal arrival rate (equivalently, the service value is high), all repeat customers with rate Λ_r join (with strictly positive utility), as well as some one-time customers (they join with zero utility). It can be shown that the total arrival rate (repeat and one-time customers) exceeds the socially optimal arrival rate. Hence,

for low potential arrival rates of repeat customers, there is over-joining. As Λ_r increases, or, alternatively, U decreases, the total arrival rate decreases until it is equal to the socially optimal arrival rate. Then, no one-time customers join. For higher values of Λ_r (or lower values of U) the one-time customers don't join anymore and the repeat customers' joining rate remains the socially optimal one. We found that:

Corollary 5.5. *In the absence of a social norm ($\kappa = 0$), the total demand rate, composed of repeat and one-time customers strictly decreases in the market size of the repeat customers as long as Λ_r is less than the socially optimal arrival rate. For higher values of Λ_r, the total demand rate is composed of repeat customers only and is remains at the socially optimal level.*

The intuition behind the decrease of the equilibrium arrival in Λ_r when both repeat and one-time customers join is the following: As the one-time customers receive the lowest priority and are served FCFS, they are more sensitive with respect to increased congestion caused by more repeat customers. Hence, one-time customers drop off at a faster rate than the rate at which repeat customers join. With insufficient repeat customers to cover the socially optimal arrival rate, the *total* joining rate is too high compared with the social welfare maximizing joining rate. Similarly to Edelson and Hildebrand (1975), this is because one-time customers, as they are served FCFS, ignore the negative externalities of their joining, causing increased waiting to other one-time customers.

With sufficient potential repeat customers, no one-time customers join, competition among repeat customers becomes intense. This tipping competition forces repeat customers to pay (via tipping) an amount that equals her net utility (service value minus waiting cost), thereby obtaining zero utility. The total arrival rate in this case (which includes only repeat customers) will be socially optimal; $\mu - \sqrt{c\mu/U}$ (as is Glazer and Hassin (1986)).

Having the equilibrium repeat customers joining rate λ and their tipping distribution at hand, we can now determine the expected tip (per customer), which is $ET(\lambda) = ET_r(\lambda)\frac{\min\{\lambda,\Lambda_r\}}{\lambda} + \underline{t}\frac{\{\Lambda_r-\lambda\}^+}{\lambda}$ where $ET_r(\lambda) = \int t dT_r(t;\lambda)$. The expected tip (and tipping distribution) is interesting from an empirical point of view. We also define the 'tipping wage' (per unit time) as $\lambda ET(\lambda)$.

The tipping wage is of managerial relevance because it is usually a credit towards the server's minimum wage. The tipping wage depends on both the tip and the equilibrium arrival rates for both customer types. As only repeat customers tip in the absence of a social norm, it is easy to obtain that $ET_r(\lambda_r) = (c/\mu)(\lambda_r/\mu)(1-\lambda_r/\mu)^{-2}$. As $\lambda_r = \min\{\mu - \sqrt{c\mu/U}, \Lambda_r\}$, we obtain that:

Corollary 5.6. *In the absence of a social norm ($\kappa = 0$), the tipping wage is $c(\Lambda_r/\mu)^2(1-\Lambda_r/\mu)^{-2}$ and increases strictly in the market size of the repeat customers, Λ_r, as long as $\Lambda_r \leq \mu - \sqrt{c\mu/U}$. For higher market sizes, the expected tipping wage achieves the maximum social welfare; $\mu U(1-\sqrt{c/(U\mu)})^2$.*

From the two Corollaries above it follows that, without a social norm, the tipping wage is non-decreasing in potential market size of repeat customers, and the total demand is non-increasing. Consequently, as long as the potential market size is relatively low, the average tip is too low compared with the optimal 'service fee', hence, there is over-joining and the average queue length, $\lambda/(\mu-\lambda)$ (an aggregate measure of service quality) is too high. When the potential market size for repeat customers is sufficiently high, the joining rate, composed of repeat customers exclusively, is socially optimal and the average tip is equal to the fee that maximizes the social welfare with a FCFS service discipline (Edelson and Hildebrand (1975)).

5.3.3 Tipping with a social norm

In this subsection, we analyze the tipping strategy in the presence of a social norm ($\kappa > 0$). In order to characterize the equilibrium tip distribution for the repeat customers, we introduce a conjecture of the unconditional expected tip, m. Given the joining rate λ and the conjectured (unconditional) expected tip m, it turns out we can characterize the repeat customers' tipping distribution as a function of m, as shown in Proposition 5.7 below. As the unconditional tipping distribution, T, is a mixture between T_r and the lowest tip in the domain, $\underline{t}$, with probabilities $\min\{\lambda, \Lambda_r\}/\lambda$ and $\{\Lambda_r - \lambda\}^+/\lambda$, respectively, we can obtain the 'actual' expected tip, for any given conjecture of the expected tip, m. Below, we define the 'rational-tipping set' as the set of conjectures, m, such that m is equal to the actual unconditional expected tip induced by this exact conjecture m.

Similarly as in §5.3.2, we subsequently characterize equilibrium arrival rates, but, for the rational tipping set. In the following, m will become an argument of all functions we already introduced. To keep notation simple, we do not differentiate the functions with (§5.3.2) and without (§5.3.3) a social norm.

For the one-time customers, the return to tipping t is a reduction in the social cost (see Equation (5.1); $-\kappa \int (\tau - t)^2 dT$). As the marginal cost of tipping t is 1, the marginal return is $-\kappa \frac{d}{dt} \int (\tau - t)^2 dT = 2\kappa(m - t)$. Note that, thanks to our quadratic social cost function, this marginal return only depends on m, *not* on the full tipping distribution T. This structure allows us obtaining analytic results: Whenever the marginal

cost of tipping is less (more) than the marginal return, the one-time customer has an incentive to increase (decrease) her tip. As tips are non-negative, the rational tipping strategy is (a degenerate distribution at) $\{m-1/(2\kappa)\}^+$. Thus, only when the expected tip is higher than $1/(2\kappa)$, one-time customers tip, otherwise, the don't (i.e., they tip zero). Proposition 5.7 fully characterizes the (non-degenerate) tipping distribution for the repeat customers and the waiting times for both customer types:

Proposition 5.7. *In the presence of a social norm ($\kappa > 0$): For a given conjecture of the expected tip amount, m, and for given $\lambda \in (0, \mu)$:*

(i) A unique equilibrium cumulative distribution function of repeat customers' tips is given by

$$T_r(t;m,\lambda) = \begin{cases} \frac{\mu}{\min\{\lambda,\Lambda_r\}} \frac{1}{\sqrt{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\mu\kappa}{c}(t^2-2(m-\frac{1}{2\kappa})t)}} + 1 - \frac{\mu}{\min\{\lambda,\Lambda_r\}}, & m \le \frac{1}{2\kappa} \\ \frac{\mu}{\min\{\lambda,\Lambda_r\}} \frac{1}{\sqrt{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\mu\kappa}{c}(t-(m-\frac{1}{2\kappa}))^2}} + 1 - \frac{\mu}{\min\{\lambda,\Lambda_r\}}, & m > \frac{1}{2\kappa} \end{cases},$$

with support $(\underline{t}(m), \bar{t}(m,\lambda)]$, such that

$$\underline{t}(m) = \{m - \frac{1}{2\kappa}\}^+,$$

and

$$\bar{t}(m,\lambda) = \begin{cases} m - \frac{1}{2\kappa} + \sqrt{(m-\frac{1}{2\kappa})^2 + \frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - 1\right\}}, & m \le \frac{1}{2\kappa} \\ m - \frac{1}{2\kappa} + \sqrt{\frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - 1\right\}}, & m > \frac{1}{2\kappa} \end{cases}$$

and is convex over its domain.

(ii) The expected waiting time as a function of the tip amount for repeat customers is:

$$W_r(t;m,\lambda) = \frac{1}{\mu(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\kappa}{c}\begin{cases} (t^2 - 2(m-\frac{1}{2\kappa})t), & m \le \frac{1}{2\kappa} \\ (t-(m-\frac{1}{2\kappa}))^2, & m > \frac{1}{2\kappa} \end{cases}$$

and for one-time customers is $W_o(\lambda)$.

The proof of Proposition 5.7 is in Appendix D.4.

Proposition 5.7 parallels Proposition 5.2. There is an additional term related to the social norm in the tipping distribution and the expected waiting time for the repeat customers. As higher tip yields higher priority, the waiting cost decreases again in the tip, t.

In the case with $\kappa = 0$, we noticed that the repeat customers' tipping distribution is convex in the tip (thus, the density is increasing). The explanation relied on the fact that the waiting cost is convex in the tipping quantile, and that the reduction in waiting cost should provide perfect compensation for the increase in the tip, such that all repeat customers receive the same utility in equilibrium. Here, with $\kappa > 0$, we have that the reduction in waiting cost should provide perfect compensation for the increase in both the tipping and the social norm costs. Note, from Equation (5.1) that the sum of the tipping and social norm costs, $t + \kappa \int (\tau - t)^2 dT)$, is convex increasing in the tip (for every conjectured mean tip m), so that the decrease in waiting cost should be even more substantial then in the case for $\kappa = 0$. Following a similar explanation as in §5.3.2, again, we have that the tipping distribution is convex in the tip, such that the density is increasing.

Being able to characterize T_r using Proposition 5.7, we can now specify the unconditional tipping distribution, T. The mean of the unconditional distribution T, $ET(m, \lambda)$, depends on the conjecture of the unconditional mean tip, m, as well as on the given joining rate λ. The expected unconditional tip can be written as:

$$ET(m, \lambda) = \frac{\min\{\lambda, \Lambda_r\}}{\lambda} ET_r(m, \lambda) + \frac{\{\lambda - \Lambda_r\}^+}{\lambda} \underline{t}(m),$$

where $ET_r(m, \lambda)$ is the (conditional) mean of the repeat customers' tipping distribution; $ET_r(m, \lambda) = \int t dT_r(t; m, \lambda)$. Next, we discuss how the expected (actual) unconditional tip, $ET(m, \lambda)$ depends on the conjecture of the tip, m:

Proposition 5.8. *For given $0 < \lambda < \mu$, we have*

(i) When $m < 1/(2\kappa)$, $ET(m, \lambda)$ is convex, strictly increasing and positive in m.

(ii) When $m > 1/(2\kappa)$, $ET(m, \lambda)$ is linear in m, and equals $ET(1/(2\kappa), \lambda) + m - 1/(2\kappa)$.

The proof of Proposition 5.8 is in Appendix D.5.

Proposition 5.8 is interesting because it prescribes the process of forming a social norm. Such norm depends on what everyone else believes is the average tip; m. When everyone else believes the expected tip is high, rational tips will also be high. According to Proposition 5.8, rational tipping with a social norm becomes more aggressive as the conjectured tip increases (ET_r is convex increasing in m). This is because the waiting time for repeat customers for a given tip t, $W_r(t; m, \lambda)$ (Proposition 5.7) is increasing in m for every t and λ. For a repeat customer tipping t, when the mean tip is m, as m increases, the same tip amount t will result in a higher waiting time. Recall that the waiting time for every tip t can be expressed as a function of merely the proportion of

tips smaller than t. This means that when m increases, every tipping quantile increases too[6]. Thus, the actual mean tip for repeat customers increases with the conjectured tip, m. Recall from the explanation in §5.3.2 that each quantile is a convex (increasing) function of the waiting time. Since $W_r(t; m, \lambda)$ is linear in $m \in (0, \frac{1}{2\kappa}]$, each quantile is convex in m, and so is ET_r (as it is a weighted sum of these quantiles).

Moreover, for sufficiently high conjectured expected tip, the tipping domain is strictly positive, i.e., $0 < \underline{t}(m)$ (see Proposition 5.7). When this is the case, an increase in the conjectured expected tip yields an equal increase in ET. In fact, notice from Proposition 5.7(ii), in case that $m > 1/(2\kappa)$, T_r, $\underline{t}$ and $\bar{t}$ are functions of $m - 1/(2\kappa)$. Hence, when $m > 1/(2\kappa)$, not only the mean of the tipping distribution increases linearly in the conjectured tip, but the entire tipping distribution is simply shifted over its domain by the same amount.

For any given joining rate λ and conjectured mean tip m, we refer to *rational tipping* when $ET = m$. That is, the distribution T, is rational when its actual mean coincides with the conjectured mean tip m. Since the expected tip is increasing in the conjectured tip, there is no guarantee that a unique rational tip exists. Define the *rational-tipping set* as:

$$\mathcal{M}(\lambda) = \{m \in \mathbb{R}^+ \mid m = ET(m, \lambda)\}.$$

The rational tipping set $\mathcal{M}$ might be an interval in $\mathbb{R}^+$. The latter is a consequence of Proposition 5.8. Define now

$$\Delta(\rho) \triangleq \frac{\frac{\sqrt{2\rho-\rho^2}}{1-\rho} + \arctan(\frac{1-\rho}{\sqrt{2\rho-\rho^2}}) - \frac{\pi}{2}}{\rho}.$$

Proposition 5.9. *For every $\kappa > 0$, let $\hat{\lambda}_r$ satisfy $\frac{1}{2\kappa} = \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu})$ and, in case that $\Lambda_r > \hat{\lambda}_r$, let $\hat{\lambda}_o$ satisfy $\frac{1}{2\kappa} = \frac{\Lambda_r}{\hat{\lambda}_o}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu})$. Then, $\hat{\lambda}_r$ decreases in the social norm, κ, $\hat{\lambda}_o > \Lambda_r$ and $\mathcal{M}(\hat{\lambda}_r) = \mathcal{M}(\hat{\lambda}_o) = [1/(2\kappa), +\infty)$.*

The proof of Proposition 5.9 is in Appendix D.6.

Proposition 5.9 identifies arrival rates for which the rational tipping set is an interval. By taking derivative with respect to ρ, it can be seen that $\Delta(\rho)$ increases from 0 to $+\infty$ as ρ increases from 0 to 1. Hence, a value $\hat{\lambda}_r \in (0, \mu)$ as defined exists uniquely. When the potential arrival rate of repeat customers is high ($\Lambda_r > \hat{\lambda}_r$), then $\hat{\lambda}_r$ is the repeat customers arrival rate at which any expected tip higher than $1/(2\kappa)$ is rational (when no one-time customers join). In that case, when μ is sufficiently large, there exists also another arrival rate, ($\hat{\lambda}_o \in [\Lambda_r, \mu)$), with one-time customers joining, at which any

[6] in other words, the tipping distribution increases stochastically with m.

expected tip higher than $1/(2\kappa)$ is rational. For any other arrival rate $\lambda \notin \{\hat{\lambda}_o, \hat{\lambda}_r\}$, the set $\mathcal{M}(\lambda)$ is either a singleton or the empty set, depending on whether the fixed point equation $m = ET(m)$ either has a unique or no solution.

Numerical Illustration: Now, we illustrate how the expected repeat customers' tip depends on the conjectured tip for various joining rates. Consider the following parameters: $\mu = 1$, $\Lambda_r = 1/4$, $c = 1$ and $\kappa = 1$. We find that the rate $\hat{\lambda}_r = 0.18384$ $(< \Lambda_r)$ solves the equation $\frac{1}{2\kappa} = \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu})$. We plot $ET(m, \lambda)$ as a function of the conjectured expected tip, m, for $\lambda = 0.9\hat{\lambda}_r$, $\hat{\lambda}_r$ and $1.1\hat{\lambda}_r$. The intersection with the 45 degree line then determines $\mathcal{M}(\lambda)$. For low joining rates of repeat customers (left panel), a unique rational tip exists and $\mathcal{M}$ is a singleton. For high joining rates (right panel), no rational tip exists ($\mathcal{M} = \emptyset$). For $\hat{\lambda}_r$, any expected tip above $1/(2\kappa) = 1/2$ is rational (middle panel).

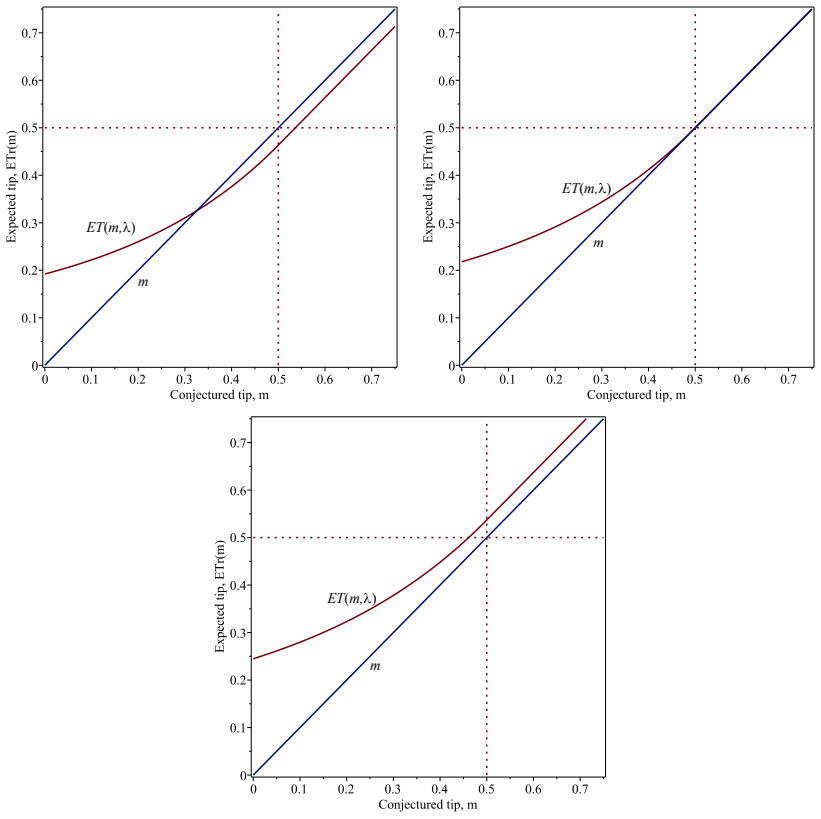


FIGURE 5.1: $ET(m, \lambda)$ as a function of m for $\lambda = 0.9\hat{\lambda}_r$, $\hat{\lambda}_r$ and $1.1\hat{\lambda}_r$ for $\mu = 1$, $\Lambda_r = 1/4$, $c = 1$ and $\kappa = 1$. This function is convex increasing with slope equals 1 for every $m \geq \frac{1}{2\kappa} = 0.5$, therefore, contingent on λ, there can exist either a unique intersection (left panel), a continuum of intersections (middle panel) or no intersection (right panel) with the 45 degree line.

Suppose that λ and m satisfy $m \in \mathcal{M}(\lambda)$. Recall that the distribution $T(t; m, \lambda)$ imposes indifference in tipping among repeat customers. In order to satisfy the equilibrium conditions, one has to verify that the customer tipping the least amount is indifferent

between joining and balking. Similarly to §5.3.2, we re-introduce $C(m;\lambda)$ as the expected total cost of a customer who tips the least amount, now, as a function of m, that is:

$$C(m;\lambda) = \underline{t}(m) + \kappa \int_{\underline{t}(m)}^{\bar{t}(m,\lambda)} (\tau - \underline{t}(m))^2 dT(\tau; m, \lambda) + \begin{cases} cW_r(\underline{t}(m); m, \lambda), & \lambda \in (0, \Lambda_r], \\ cW_o(\lambda), & \lambda \in (\Lambda_r, \mu). \end{cases}$$

Finally, as we are interested in equilibria with rational tipping, for a given joining rate $\lambda < \mu$, we define the set of *rational costs*, $\mathcal{C}(\lambda)$, that is, the image of $C(m;\lambda)$ as a function of m on the rational tipping set, $\mathcal{M}(\lambda)$:

$$\mathcal{C}(\lambda) = \{C \in \mathbb{R}^+ \mid C = C(m;\lambda), m \in \mathcal{M}(\lambda)\}.$$

In the case that $\mathcal{M}(\lambda) = \emptyset$, by definition, also $\mathcal{C}(\lambda) = \emptyset$. A value $\lambda < \mu$ thus prescribes equilibrium joining rate (Equation (5.2)) if (and only if):

$$U \in \mathcal{C}(\lambda).$$

With a slight abuse of notation, for any arrival rate $\lambda < \mu$ such that $\mathcal{C}(\lambda) \neq \emptyset$ we define, with one argument only:

$$C(\lambda) = \min\{C | C \in \mathcal{C}(\lambda)\},$$

which is the value in $\mathcal{C}(\lambda)$ when the latter is a singleton, or, the lowest cost in $\mathcal{C}(\lambda)$ when it is an interval. In the former case, the corresponding rational tip is the unique element of $\mathcal{M}(\lambda)$. In the latter case, the cost minimizing tip is the lowest one; $1/(2\kappa)$. Next, we illustrate $\mathcal{C}(\lambda)$ numerically.

Numerical Illustration: Consider the following parameters: $\mu = 1$, $\Lambda_r = 1/4$, $c = 1$ and $\kappa = 1$. Figure 5.2 below illustrates $\mathcal{C}(\lambda)$. The left side depicts the rational cost set $\mathcal{C}(\lambda)$ for every value $\lambda \in [0, \mu)$, where $\hat{\lambda}_r = 0.1838$. Notice that $\Lambda_r = 1/4 > \hat{\lambda}_r$ and therefore, $\hat{\lambda}_o$ exists; $\hat{\lambda}_o = 0.318$.

It is noteworthy that no costs are defined over $(\hat{\lambda}_o, \hat{\lambda}_r)$ as the rational tipping set, $\mathcal{M}$ is empty for these arrival rates. Furthermore, for every $\lambda < \hat{\lambda}_r$ (no one-time customers joining), the rational tipping set $\mathcal{M}(\lambda)$ is a singleton, thus, so is $\mathcal{C}(\lambda)$, and in fact, the unique value associated with each such λ in this case is $C(\lambda)$. However, when $\lambda = \hat{\lambda}_r$, the rational tipping set $\mathcal{M}(\hat{\lambda}_r) = [1/(2\kappa), +\infty)$ is an unbounded interval, and so is $\mathcal{C}(\hat{\lambda}_r)$. Similarly, any $\lambda > \hat{\lambda}_o$ implies that both $\mathcal{C}(\lambda)$ and $\mathcal{M}(\lambda)$ are singletons, and associated with each such λ is the unique value $C(\lambda)$ on the vertical axis. For $\lambda = \hat{\lambda}_o$ again we have that $\mathcal{C}(\lambda)$ is an unbounded interval and so, a rational cost at $\hat{\lambda}_o$ can be arbitrarily large.

In the left panel, $C(\lambda)$ is strictly increasing and bounded as a function of λ over the interval $[0, \hat{\lambda}_r)$. However, considering $C(\lambda)$ over the interval $(\hat{\lambda}_o, \mu)$, it is non-monotone, and grows arbitrarily large when λ approaches μ. In Lemmas 5.10 and 5.12 below, we formalize the behavior of $\mathcal{C}(\lambda)$, as these determine the structure of the equilibrium.

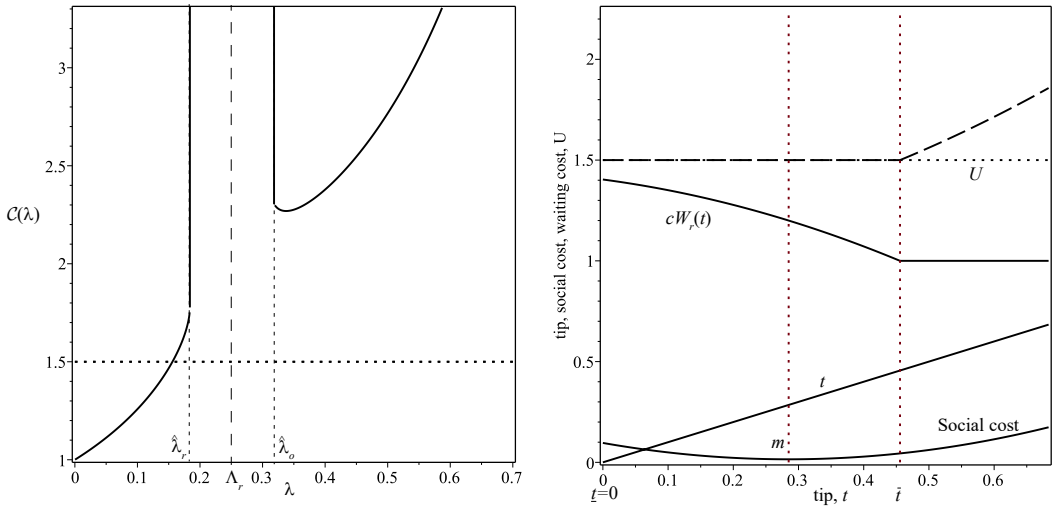


FIGURE 5.2: For $\mu = 1$, $\Lambda_r = 1/4$, $c = 1$ and $\kappa = 1$, the left panel illustrates $\mathcal{C}(\lambda)$. Note that $\mathcal{C}(\lambda)$ is a singleton and increasing over $\lambda \in (0, \hat{\lambda}_r)$, an interval for $\lambda = \hat{\lambda}_r$ or $\lambda = \hat{\lambda}_o$, an empty set for $\lambda \in (\hat{\lambda}_r, \hat{\lambda}_o)$ and a singleton and convex over $\lambda \in (\hat{\lambda}_o, \mu)$. The right panel depicts the decomposition of the equilibrium total cost and U. The three components of the cost – tipping, waiting and social norm are in solid lines. The total cost is indicated with a dashed line, which is equal to U for all $t \in (0, \bar{t}]$ and strictly higher for $t > \bar{t}$.

For $U = 3/2$ (horizontal dotted line in both panels), there exists exactly one equilibrium with only repeat customers joining; $3/2 \in \mathcal{C}(\lambda)$ for $\lambda = 0.15594 < \hat{\lambda}_r$. As $\mathcal{C}(\lambda)$ is a singleton with a unique rational tipping strategy, we obtain, for this value λ, that $m = 0.2849$, with $\underline{t} = 0$ and $\bar{t} = 0.4556$. In the right panel, we plot the repeat customer's costs (tipping, waiting and social norm cost) as a function of their tip amount. Over the randomization domain the sum of these three costs is independent of the tip and is equal to U. This makes repeat customers indifferent between joining and not and supports randomization of the joining decision over the tipping domain. Outside the randomization domain the sum of these three costs is strictly higher than inside; repeat customers prefer not to tip any such amount. From Proposition 5.1, it follows that the one-time customers strictly prefer not joining. Notice that the social norm cost is a quadratic function centered inside the randomization domain at the expected tip level (m). As $m < 1/(2\kappa) = 1/2$, the joining one-time customers don't tip $(\underline{t} = 0)$.

Finally, notice that for higher service values, multiple equilibria may exist. For example, when $U = 3$, exactly three equilibria exist: $3 \in \mathcal{C}(\hat{\lambda}_r)$ induces an equilibrium with repeat customers only joining at rate $\hat{\lambda}_r$; $3 \in \mathcal{C}(\hat{\lambda}_o)$ induces an equilibrium with repeat and one-time customers joining at rates Λ_r and $\hat{\lambda}_o - \Lambda_r$ respectively; There exists $\lambda \in (\hat{\lambda}_o, \mu)$ such that $3 = C(\lambda) \in \mathcal{C}(\lambda)$ inducing an equilibrium with repeat and one-time customers joining at rates Λ_r and $\lambda - \Lambda_r$ respectively.

In the next subsection, we first focus on equilibria without one-time customers, i.e., $\lambda \leq \Lambda_r$. Then, on equilibria with one-time customers, i.e, $\lambda > \Lambda_r$.

Equilibria with repeat customers only ($\lambda \in (0, \Lambda_r]$)

The following Lemma confirms the shape of $\mathcal{C}(\lambda)$, $\lambda \in (0, \Lambda_r]$ from Figure 5.2 (left side) in the numerical illustration:

Lemma 5.10. *For every $\lambda \in (0, \Lambda_r)$: When $\hat{\lambda}_r < \Lambda_r$, then, for every $\lambda \in (0, \Lambda_r]$:*

(i) If $\lambda < \hat{\lambda}_r$, then $\mathcal{C}(\lambda)$ is a singleton, and increasing in λ.

(ii) If $\lambda = \hat{\lambda}_r$, then $\mathcal{C}(\lambda) = [\hat{C}_r, +\infty)$ is an unbounded interval, where $\hat{C}_r = C(\hat{\lambda}_r)$.

(iii) If $\lambda > \hat{\lambda}_r$, then $\mathcal{C}(\lambda) = \emptyset$.

When $\Lambda_r < \hat{\lambda}_r$, then $\mathcal{C}(\lambda)$ is a singleton, and increasing in $\lambda \in (0, \Lambda_r]$.

The proof of Lemma 5.10 is in Appendix D.7.

The Proposition below characterizes the structure of the equilibria with repeat customers only ($\lambda < \Lambda_r$), which we elaborate next:

Proposition 5.11. *For any $\kappa > 0$, c, and μ*

(i) When $\Lambda_r < \hat{\lambda}_r$, then:

(i-a) If $U > C(\Lambda_r)$, there exist no equilibria with only repeat customers.

(i-b) If $U \leq C(\Lambda_r)$, there exists an equilibrium with only repeat customers and with mean tip $m < 1/(2\kappa)$.

(ii) When $\Lambda_r \geq \hat{\lambda}_r$, then:

(ii-a) If $U \geq \hat{C}_r$, there exists an equilibrium with only repeat customers and with mean tip $m = 1/(2\kappa) + U - \hat{C}_r$.

(ii-b) If $U < \hat{C}_r$, there exists an equilibrium with only repeat customers and with mean tip $m < 1/(2\kappa)$.

The proof of Proposition 5.11 is in Appendix D.8.

Proposition 5.11(i) implies that for $\Lambda_r < \hat{\lambda}_r$, equilibria with repeat customers only exist when the service value is low, otherwise, as we will see in Proposition 5.14, one-time customers join too in equilibrium. Proposition 5.11(ii) implies that if $\Lambda_r > \hat{\lambda}_r$, not all repeat customers can join in equilibrium. The term $\hat{\lambda}_r$ caps the arrival rate of repeat customers. $\hat{\lambda}_r$ depends on the social norm, service capacity and waiting cost, but neither depends on the potential arrival rate of repeat customers, nor on the service value (similarly to the case without a social norm). This cap is an important feature of our model.

The intuition behinds the existence of such cap $\hat{\lambda}_r$ is the following: Assume that all potential repeat customers join. When the potential demand is relatively high, for every belief m of mean tip amount in the population, obtaining faster service is so valuable that a customer wants to deviate by tipping more. As a consequence, the best response of all other customers is to tip even more and hence, the total cost spirals out of control. Therefore, if all the repeat customers join with rate $\Lambda_r > \hat{\lambda}_r$, no mean tip satisfies rational tipping (i.e. $\mathcal{M} = \emptyset$). Hence, an equilibrium can only be reached making tipping less competitive through alleviating congestion. Thus, not all repeat customers can join. The only joining rate that makes this possible for $U > \hat{C}_r$ is $\hat{\lambda}_r < \Lambda_r$.

From Proposition 5.8, it follows that any expected tip exceeding $1/(2\kappa)$ is rational and shifts the actual mean tip with the amount exceeding $1/(2\kappa)$. With repeat customers joining at rate $\hat{\lambda}_r$ and a conjectured tip of $1/(2\kappa)$, the cost is $\hat{C}_r$. Hence, the shift in expected tip that makes repeat customers indifferent between joining and balking is determined by $m - 1/(2\kappa) + \hat{C}_r = U$.

Thus, the social norm ($\kappa > 0$) introduces a 'tipping' war among all repeat customers that lifts up the expected tip, making the cost of joining (including tipping, waiting and social norm cost) equal to the service value. Such tipping war is driven by the fundamental difference of customer interaction with each other in stage one of our model ('service consumption') versus stage two of our model ('social consumption'). The service interaction occurs in a capacitated environment in which the value of tipping (reducing wait cost) drives up the tip. In the social interaction ensures that the tip is not too high. When the service value, U, is sufficiently high ($U > \hat{C}_r$), the former effect dominates, making the repeat customer's lowest tip strictly positive; $\underline{t} > 0$. When the U is low, the lowest tip is zero. This is different from Proposition 5.4, without a social norm, for which the lowest tip was *always* zero.

Equilibria with repeat customers and one-time customers ($\lambda \in (\Lambda_r, \mu)$)

The following Lemma 5.12 confirms the shape of $\mathcal{C}(\lambda)$, $\lambda \in (\Lambda_r, \mu)$ from Figure 5.2 (left side panel) in the numerical example.

Lemma 5.12. *Define $\hat{\Lambda}_r$ such that $\frac{1}{2\kappa} = \frac{\hat{\Lambda}_r}{\mu}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\Lambda}_r}{\mu})$, we have that $\hat{\lambda}_r < \hat{\Lambda}_r$. When $\hat{\lambda}_r < \Lambda_r < \hat{\Lambda}_r$, then, for every $\lambda \in (\Lambda_r, \mu)$:*

(i) If $\lambda < \hat{\lambda}_o$, then $\mathcal{C}(\lambda) = \emptyset$.

(ii) If $\lambda = \hat{\lambda}_o$, then $\mathcal{C}(\lambda) = [\hat{C}_o, +\infty)$ is an unbounded interval, where $\hat{C}_o = C(\hat{\lambda}_o)$

(iii) If $\lambda > \hat{\lambda}_o$, then $\mathcal{C}(\lambda)$ is a singleton, and convex in λ.

When $\Lambda_r < \hat{\lambda}_r$, then $\mathcal{C}(\lambda)$ is a singleton, and convex in $\lambda \in (\Lambda_r, \mu)$.

When $\hat{\Lambda}_r < \Lambda_r$, then $\mathcal{C}(\lambda) = \emptyset$ for $\lambda \in (\Lambda_r, \mu)$.

The proof of Lemma 5.12 is in Appendix D.9.

When the one-time customers join, they exert both negative and positive externalities on other one-time customers: On the one hand, they increase the waiting time for other joining one-time customers in the consumption stage. On the other hand, they make it more likely meeting up with another one-time customer in the social market stage. Such encounter leads to zero dis-utility as all one-time customers tip alike (Proposition 5.1). As a consequence, it is possible that their total cost decreases in their arrival rate $\lambda - \Lambda_r$.

In addition, the one-time customers exert positive externalities on the repeat customers too: the latter's total cost decreases with $\lambda > \Lambda_r$: Recall from Proposition 5.1 that one-time customers tip the lowest among all customers, thus, when they join, they drag down the mean tip. This leads to reduction in the repeat customer's tip cost, and, as the repeat customers' waiting cost is invariant to the joining of one-time customers (because of the preemptive priority), repeat customers enjoy a reduction in the total cost.

From a queuing perspective, it is interesting that a higher arrival rate can result in less costs. In our model, a possible increase in waiting costs may be more than offset by a decrease in tipping costs.

Suppose that $\Lambda_r > \hat{\lambda}_r$. Recall that Proposition 5.11 suggests that in equilibrium without one-time customers ($\lambda < \Lambda_r$) the largest possible joining rate of repeat customers is $\hat{\lambda}_r$. This is why in Figure 5.2, the rational cost set is empty for arrival rates in $(\hat{\lambda}_r, \Lambda_r)$.

When one-time customers join too, more than $\hat{\lambda}_r$ repeat customers can join due to the aforementioned reduction of cost. In fact, according to Lemma 5.12, when the arrival rate of one-time customers is sufficiently high (such that the total arrival rate is $\hat{\lambda}_o$), *all* repeat customers (with rate Λ_r) can join. Hence, a critical, minimal volume of one-time customers ($\hat{\lambda}_o - \Lambda_r$) is required to reduce the repeat customers' costs sufficiently to entice them to join again. This is why in Figure 5.2, the rational cost set is also empty for arrival rates in $(\Lambda_r, \hat{\lambda}_o)$. Yet, when Λ_r is high, since the total capacity of the system is limited to μ, it is possible that there is not enough service capacity left for one-time customers to allow all repeat customers to join with positive expected utility. Thus, $\hat{\Lambda}_r$ is the maximal potential rate of repeat customers, such that even if they all join, there is still enough capacity left for one-time customers to allow rational tipping.

Recall that when $\Lambda_r > \hat{\lambda}_r$, e.g. when the social norm is sufficiently strong (see Proposition 5.9), the repeat customers wage a tipping war among themselves. In that case, we have that $\hat{\lambda}_o > \Lambda_r$ and, with sufficient amount of one-time customers to reduce the mean tip, a tipping war among *all* customers is possible.

Define

$$\underline{C} = \min_{\lambda \in [\Lambda_r, \mu)} C(\lambda),$$

which is the lowest rational cost with one-time customers joining ($\lambda \in [\Lambda_r, \mu)$), and denote its minimizer by $\underline{\lambda}$.[7]

Lemma 5.13. *For any $\kappa > 0$, c, and μ, there exist $\underline{\Lambda}'_r \leq \underline{\Lambda}''_r$ and $\bar{\Lambda}_r$ satisfying $0 < \underline{\Lambda}'_r \leq \underline{\Lambda}''_r < \hat{\lambda}_r < \bar{\Lambda}_r < \hat{\Lambda}_r$, such that:*

(i) If $\Lambda_r \in (\underline{\Lambda}''_r, \bar{\Lambda}_r)$, then $C(\Lambda_r) > \underline{C}$ for $\Lambda_r < \hat{\lambda}_r$ and $\hat{C}_o > \underline{C}$ for $\hat{\lambda}_r < \Lambda_r$;

(ii) If $\Lambda_r \notin (\underline{\Lambda}'_r, \bar{\Lambda}_r)$, then $C(\Lambda_r) = \underline{C}$ for $\Lambda_r < \hat{\lambda}_r$ and $\hat{C}_o = \underline{C}$ for $\hat{\lambda}_r < \Lambda_r$.

The proof of Lemma 5.13 is in Appendix D.10.

Through extensive numerical experimentation, we found that $\underline{\Lambda}'_r = \underline{\Lambda}''_r$, even though a formal proof of it escapes us. In essence, according to Lemma 5.13, only when the potential arrival rate for repeat customers is in the neighborhood of $\hat{\lambda}_r$ (either higher or lower), the positive externalities that the one-time customers exert for arrival rates higher than $\hat{\lambda}_o$ dominate the negative externalities. When Λ_r is less (more) than $\hat{\lambda}_r$, the lowest arrival rate with one-time customers is Λ_r ($\hat{\lambda}_o$). In these cases, the costs $C(\lambda)$ decreases first and then increases. Per our equilibrium condition, $U \in \mathcal{C}(\lambda) = C(\lambda)$,

[7] From Lemma 5.12, $C(\lambda)$ is convex over $[\hat{\lambda}_o, \mu)$, thus $\underline{\lambda}$ exists uniquely.

two equilibria are possible for $U \in (\underline{C}, \hat{C}_o]$, one below $\underline{\lambda}$ and one above. The following proposition provides a full overview of all possible equilibria[8]:

Proposition 5.14. *For any $\kappa > 0$, c, and μ*

(i) When $\Lambda_r \leq \hat{\lambda}_r$, then:

(i-a) If $U > C(\Lambda_r)$, there exists one equilibrium with repeat and one-time customers for some $\lambda > \underline{\lambda}$ with mean tip $m \leq 1/(2\kappa)$.

(i-b) If $U \in (\underline{C}, C(\Lambda_r)]$, there exist two equilibria with repeat and one-time customers:
λ' for some $\lambda' \in [\hat{\lambda}_o, \underline{\lambda})$ with mean tip $m' \leq 1/(2\kappa)$;
λ'' for some $\lambda'' \in (\underline{\lambda}, \mu)$ with mean tip $m'' \leq 1/(2\kappa)$.

(i-c) If $U < \underline{C}$, there exist no equilibria with repeat and one-time customers.

(ii) When $\Lambda_r \in (\hat{\lambda}_r, \hat{\Lambda}_r)$, then:

(ii-a) If $U > \hat{C}_o$, then there exist two equilibria with repeat and one-time customers:
$\hat{\lambda}_o$ with mean tip $m = 1/(2\kappa) + U - \hat{C}_o$; and
λ for some $\lambda > \underline{\lambda}$ with mean tip $m \leq 1/(2\kappa)$.

items (ii-b), (ii-c), are identical to (i-b), (i-c), respectively but with $\hat{C}_o$ instead of $C(\Lambda_r)$.

(iii) When $\Lambda_r \geq \hat{\Lambda}_r$, there exist no equilibria with repeat and one-time customers.

The proof of Proposition 5.14 is in Appendix D.11.

In an equilibrium where the mean tip, m, is not greater than $1/(2\kappa)$, we have that the lowest tip $\underline{t}$ (that is, the tip paid by one-time customers) is zero (see Proposition 5.7). According to Proposition 5.14(i), all equilibria for low arrival rates of repeat customers ($\Lambda_r \leq \hat{\lambda}_r$), or for low service values ($U \leq \hat{C}_o$), have zero-tipping one-time customers. Thus, even with a social norm ($\kappa > 0$), there is not necessarily a reason for one-time customers to tip. When one-time customers don't tip, there is an equilibrium for high enough service value (i-a) and none for low service value (i-c). For intermediate service value (i-b), there might be two equilibria for potential arrival rates of repeat customers in the neighborhood of $\hat{\lambda}_r$, as discussed in Lemma 5.13. Obviously, for too high potential arrival rates of repeat customers (Proposition 5.14(iii)) no equilibria with both one-time and repeat customers exist.

[8] For sake of expositional clarity, we do not report knife-edge cases ($U = \underline{C}$). The latter can simply be obtained by continuity.

Interestingly, according to Proposition 5.14(ii) for medium arrival rates of repeat customers, when $\Lambda_r \in (\hat{\lambda}_r, \hat{\Lambda}_r)$, and high enough service value, $(U > \hat{C}_o)$, there exists an equilibrium in which one-time customers tip; $\underline{t} > 0$:

Corollary 5.15. *In the presence of a social norm, only when the potential arrival rate of repeat customers is intermediate and the service value is high ($\Lambda_r \in (\hat{\lambda}_r, \hat{\Lambda}_r)$ and $U > \hat{C}_o$), the one-time customers join and their tip is strictly positive.*

The Corollary provides an answer to the question of when a social norm motivates one-time customers to tip. For this equilibrium (Proposition 5.14(ii-a)), we have that $\hat{\lambda}_o > \Lambda_r$ and therefore, an *all out tipping war* rages among a mixture of one-time and repeat customers. As in the tipping war with repeat customers only (at arrival rate $\hat{\lambda}_r$), here, the congestion effects cause by finite capacity make tipping valuable again, causing the rational tip to spiral out of control; $\mathcal{M}(\hat{\lambda}_o) = [1/(2\kappa), +\infty)$. This equilibrium exists thanks to the critical, minimum volume of one-time customers $(\hat{\lambda}_o - \Lambda_r)$ whose presence puts downward pressure on the tip cost, allowing all repeat customers to join at a rate $\Lambda_r (> \hat{\lambda}_r)$. In our discussion below, this equilibrium will play an important role as, for some parameter values, the tipping wage will be the highest.

Overview of equilibria

Now, for $c = 1$, $\kappa = 1$ and $\mu = 1$ we numerically illustrate all equilibria identified above as a function of the potential arrival rate of repeat customers, Λ_r, and the service value, U. Note that for $\Lambda_r < \hat{\lambda}_r$, if $U \leq C(\Lambda_r)$, then by Proposition 5.11 item (i-b) there exists an equilibrium with only repeat customers joining, and if $U > C(\Lambda_r)$, then by Proposition 5.14 item (i-b) there exists an equilibrium with both repeat and one-time customers joining. Hence, $C(\Lambda_r)$ partitions the (Λ_r, U)-plane into two regions. If on the contrary $\Lambda_r \geq \hat{\lambda}_r$, then, as $\hat{\lambda}_r < \hat{\lambda}_o$, we have that

$$\hat{C}_r = C(1/(2\kappa), \hat{\lambda}_r) \leq C(1/(2\kappa), \hat{\lambda}_o) = \hat{C}_o.$$

These two curves partition the (Λ_r, U) space in three regions. Finally, recall from Lemma 5.13 that the regions around $\hat{\lambda}_r$ for which $\underline{C}$ (indicated by dashed lines) is strictly less than $C(\Lambda_r)$ or $\hat{C}_o$ introduce another region in which we have two equilibria. We plot $C(\Lambda_r)$, $\hat{C}_r$, $\hat{C}_o$ and $\underline{C}$ as functions of Λ_r. Notice that the regions determined by $\underline{C}$ are relatively small and hence, to keep focus, we do not discuss the equilibria (Proposition 5.14(ii-a)) in these regions.

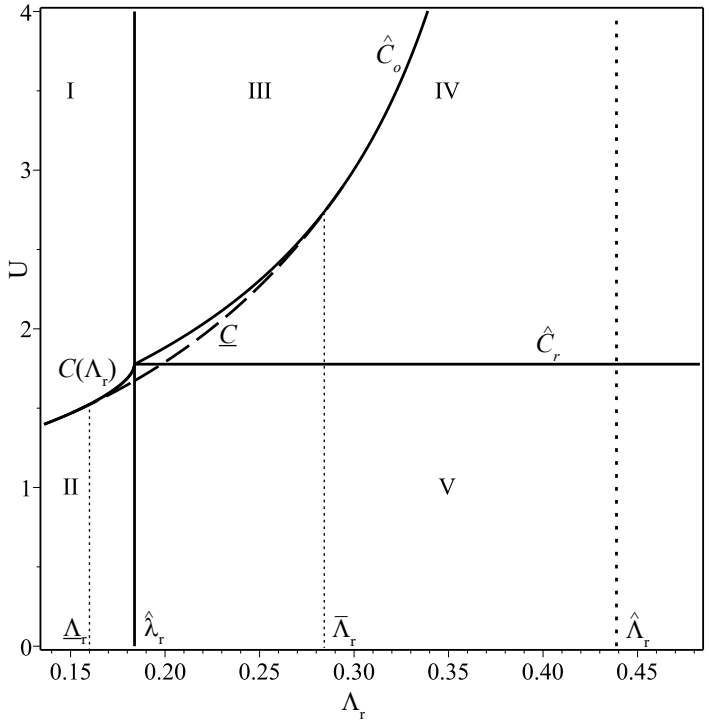


FIGURE 5.3: Overview of equilibria for $\mu = 1$, $\kappa = 1$ and $c = 1$. The (Λ_r, U)-plane is devided by the curves $C(\Lambda_r)$, $\hat{C}_r$, $\hat{C}_o$ and $\underline{C}$ into several partitions, among which we focus on five, indicated by I-V: In Regions I, III, equilibria exists for which one-time as well as repeat customers join. In Regions II, III, IV and V, equilibria exists for which only repeat customers join. Region III is such that an equilibria exists with one-time customers tip being strictly positive.

Insights

Equilibrium tipping wage and demand

Recall that $\hat{\lambda}_r$ depends only on the wait cost, c, service rate, μ and social norm cost, κ. To illustrate sharply, consider the knife-edge parameters for which $\Lambda_r = \hat{\lambda}_r$ (and some $\kappa > 0$). From Proposition 5.11(ii-a), we know that for high enough service value, $U > \hat{C}_r$, there exists an equilibrium with repeat customers joining exclusively with rate $\hat{\lambda}_r$, and mean tip $U - \hat{C}_r + 1/(2\kappa) \in \mathcal{M}(\hat{\lambda}_r) = [1/(2\kappa), +\infty)$. For this equilibrium, as no one-time customers join, rational tipping is determined by repeat customers only. Since $\Lambda_r = \hat{\lambda}_r$, we have $\hat{\lambda}_o = \Lambda_r$, thus, by definition, $\hat{C}_r = \hat{C}_o$, therefore $U > \hat{C}_o$. From Proposition 5.14(ii-a), there exists also an equilibrium with one-time customers joining; $\lambda > \hat{\lambda}_r$, but, they don't tip ($\underline{t} = 0$). This is always possible because encounters in the social market stage with non-tipping one-time customers reduces the expected tip for the one-time customers; with $\lambda - \Lambda_r > 0$, there exists a rational tip $m \in (0, 1/(2\kappa))$, i.e. with one-time customers joining, but, not tipping. The presence of non-tipping one-time customers in the second equilibrium drags down the mean tip, but, increases the total volume of customers (from $\hat{\lambda}_r$ to $\lambda > \hat{\lambda}_r$). The tipping wage for the equilibrium

with one-time customers is therefore $\lambda \cdot m = \hat{\lambda}_r ET_r(m, \hat{\lambda}_r)$, while the tipping wage for the equilibrium with repeat customers exclusively is $(U - \hat{C}_r + 1/(2\kappa))\hat{\lambda}_r$. Now, we can compare the tipping wage with and without one-time customers. Note that $1/(2\kappa) = ET_r(1/(2\kappa), \hat{\lambda}_r) > ET_r(m, \hat{\lambda}_r)$ (as $1/(2\kappa) > m$). Therefore, as $U > \hat{C}_r$, the tipping wage in the equilibrium without one-time customers, $(U - \hat{C}_r + 1/(2\kappa))\hat{\lambda}_r$, is higher than in the equilibrium with one-time customers $ET_r(m, \hat{\lambda}_r)\hat{\lambda}_r$:

Corollary 5.16. *In the presence of a social norm, when $\Lambda_r = \hat{\lambda}_r$, there exists one equilibrium without one-time customers joining and one equilibrium with one-time customers joining, but, they don't tip. The tipping wage in the former equilibrium is higher than in the latter. The demand in the former equilibrium, however, is lower than in the latter.*

The Corollary illustrates the existence of multiple equilibria in which tipping wage and customer demand are traded off; a service facility attracting only repeat customers, who rationally tip high for fast service and therefore 'tip out' the one-time customers or a service facility attracting a mixture of repeat and non-tipping one-time customers.

Numerical Illustration: Figure 5.3 illustrates the expected tipping wage as a function of the market size of repeat customers for $c = 1$, $\mu = 1$, $\kappa = 1$ and $U = 9$. For intermediate market size of repeat customers ($\Lambda_r > \hat{\lambda}_r$), there exist multiple equilibria. Recall from Corollary 5.5 that the tipping wage (total demand) increases (decreases) monotonically in the potential market demand of repeat customers and stays flat above the socially optimal arrival rate. With a social norm, the figure shows a different behavior, in part due to the multiplicity of equilibria. We already argued that, with a social norm, at Λ_r equal to $\hat{\lambda}_r$, the equilibrium with exclusive repeat customers has a higher tipping wage than the one with non-tipping one-time customers (but, lower demand). For Λ_r slightly higher than $\hat{\lambda}_r$, there is another, third, equilibrium in which the one-time customers join and tip a strictly positive amount (Proposition 5.14(i-a)). This equilibrium achieves, at some intermediate Λ_r, the highest possible tipping wage at an intermediate total demand rate.

It is interesting to note from the right panel that for the equilibrium arrival rate can be either above or below the socially optimal arrival rate (which is $2/3 \approx 0.666$ for the parameter values in Figure 5.4, right panel). As the average queue length, which is a measure of 'service quality', depends on the arrival rate only (and is equal to $\lambda/(\mu - \lambda)$), the latter implies that in equilibrium, service quality can be over- or under-provided. Recall that without a social norm ($\kappa = 0$), the average queue length was too long when the potential market of repeat customers is too small and hence, quality is under-provided in equilibrium. In Azar (2005b), the service quality is under-provided in equilibrium

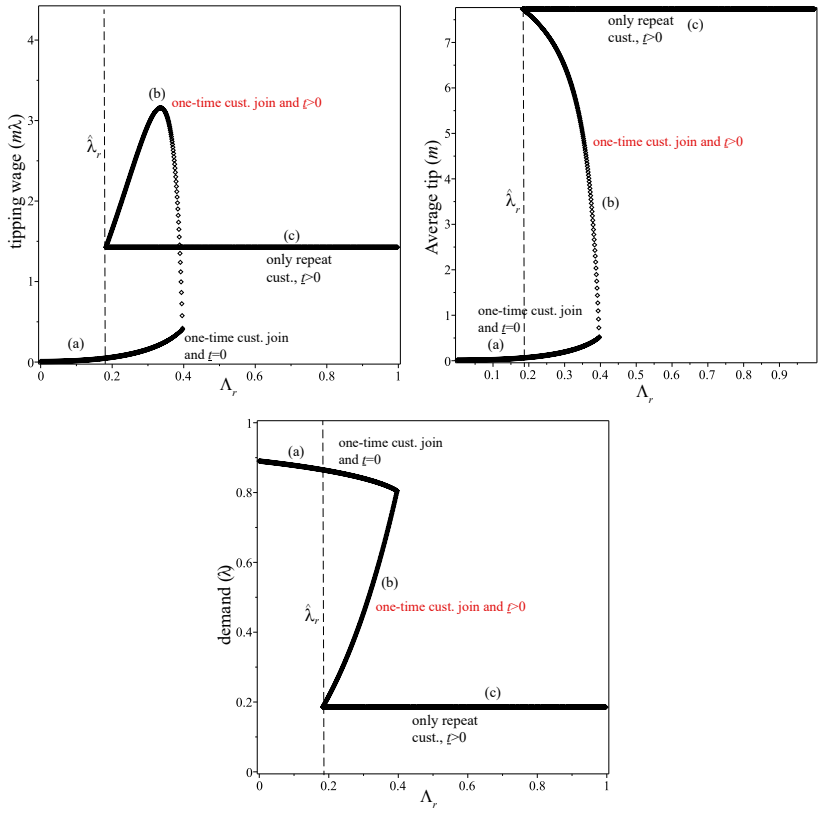


FIGURE 5.4: Expected tipping wage for $\mu = 1$, $\kappa = 1$ and $c = 1$ and $U = 9$. Curve (a) are equilibria with one-time customers joining, but, not tipping ($\lambda > \Lambda_r, m < 1/(2\kappa)$, Regions I & III in Figure 5.3). Curve (b) are equilibria with one-time customers joining and tipping ($\lambda > \Lambda_r, m = 1/(2\kappa) + U - \hat{C}_o$, Region III in Figure 5.3). Curve (c) are equilibria with no one-time customers joining ($\lambda > \hat{\lambda}_r, m = 1/(2\kappa) + U - \hat{C}_r$, Regions III & IV in Figure 5.3).

(under reasonable parameter settings). In our model, there exist also equilibria in which there is also over-provision of service quality. Intuitively, this is because the tipping war restricts the joining rate, which improves the quality.

Social welfare and tipping with a social norm.

Now, we study the role of the social norm and tipping on the social welfare. The maximum social welfare that can be achieved is $\mu U(1 - \sqrt{c/(U\mu)})^2$ (see e.g. Glazer and Hassin (1986)). Without a social norm ($\kappa = 0$), we already discussed that the tipping wage can achieve the maximum social welfare, but only when there are sufficient repeat customers, such that no one-time customers join (Corollary 5.6). When one-time customers do join, because they are served FCFS, their existence in the system creates market inefficiencies due to over-joining of the facility.

To the extent that $\kappa > 0$, the existence of a social norm carries out an inherent loss in welfare: In principle, customers could all tip exactly the same amount and avoid any costs in the social market stage. However, this is impossible in congested markets with

repeat customers, as explained in Proposition 5.1; repeat customers have incentives to differentiate their tips, thanks to the priority service discipline. By model design, any strictly positive social norm, $\kappa > 0$, incurs strictly positive costs bore by the customers. Hence, in the presence of a social norm, the maximum welfare is never achieved in our model.

Nevertheless, it is possible that social welfare losses are minimal. Notice from Figure 5.4 that for an interval of arrival rates Λ_r just above $\hat{\lambda}_r$, the tipping wage increases in Λ_r, attains a maximum and then decreases. Recall from Proposition 5.9 that $\hat{\lambda}_r$ decreases in the social norm, κ. As the social norm gets stronger (i.e., κ increases), $\hat{\lambda}_r$ approaches zero. Corollary 5.17 characterizes a relation between κ and Λ_r, in which, when κ is large, the interval over which the tipping wage increases and then decreases shifts towards values of Λ_r close to zero (the order of $1/\kappa^{1/3}$). Moreover, the maximum tipping wage over this interval approaches $\mu U(1-\sqrt{c/(U\mu)})^2$, i.e., the maximum social welfare. Contrary to a weak social norm, the welfare losses are minimal when there are few repeat customers, who all join, as well as one-time customers, who all tip a strictly positive amount. As the one-time customers join, the repeat customers obtain strictly positive surplus, hence, to achieve minimal welfare losses, their volume must be vanishing small. Finally, from Proposition 5.7, observe that the tipping distribution of these repeat customers will be concentrated around the mean tip ($\underline{t} \to \bar{t}$), which is necessary to minimize the social cost, caused by the variance of the tipping distribution.

Corollary 5.17. *(i) In the presence of a social norm ($\kappa > 0$), the tipping wage is always strictly less than the optimal social welfare for any $\Lambda_r \in (0, \mu)$.*

(ii) In the presence of a strong social norm ($\kappa \gg 0$), when the arrival rate of repeat customers is low ($\Lambda_r = \frac{1}{2}\{\frac{3\mu^2}{2\sqrt{\kappa c}}(1-\sqrt{c/(U\mu)})\}^{\frac{2}{3}}$), the joining rate is approximately socially optimal, one-time customers join and tip, and the social welfare losses become arbitrary small.

Technical derivation of Corollary 5.17(ii) is given in the Appendix.

Our model illustrates the sources of the losses in social welfare in the tipping mechanism: (a) the strictly positive utility repeat customers enjoy and (b) the strictly positive social costs, caused by the repeat customers' tipping behavior. Without social norm, there must be enough repeat customers in the market such that they competitively tip to eliminate (a). With a strong social norm the welfare losses are minimized when the tip distribution is concentrated, to reduce (b) *and* there are not many repeat customers, to reduce (a).

The above finding addresses an equilibrium where the market of repeat customers is sufficiently small, and the significant share of the capacity is allocated by one-time

customers. Note that in order to make them tip, the all-out tipping war is necessary and, as discussed above, such war is driven by the finite service capacity. In such case, the system resembles an FCFS queue, in which the *variance* of the waiting time is minimized[9]. In our model, waiting time can be thought of as a measure of 'service quality'. Empirically, the relationship between 'service quality' and tips is ambiguous, May (1978), Lynn and Grassman (1990). Our model sheds a light on this: a stronger social norm does indeed reduce the tip and service quality variability. In the Appendix, we explore numerically how tip and service quality (i.e. waiting time) co-vary. We focus on the equilibrium in which one-time customers join and tip. Even though the standard deviation of the tip can increase in the social norm, we observe that for a very strong social norm, the standard deviation decreases; i.e. the tipping distribution becomes more concentrated. Interestingly, the standard deviation of the waiting time increases as the social norm becomes stronger. This is because of the demand effect in our model; in the equilibrium at which one-time customers tip, more customers join when the social norm is stronger, creating more congestion and therefore more waiting time variability. As expected, a higher tip implies a lower waiting time and therefore, we observe negative correlation between the two. We also observe that as the social norm becomes stronger, the correlation decreases. This implies that tips become better predictors of waiting time (service quality). With the demand (joining rate) determined endogeneously and the limited service capacity in our model, a stronger social norm amplifies the role of tips on service quality, due to increased congestion.

Finally, we discuss the social welfare in an extension of our model. Instead of incurring dis-utility (costs) in the social consumption phase, customers could obtain strictly positive utility, say ν, from tipping d more than the randomly selected customer with whom she is paired. Such preferences can be justified e.g. when over-tippers ($d > 0$) are considered to be generous and the act of tipping generates positive feelings ($\nu > 0$), see Azar (2004b, 2005b). Instead of $\kappa(\tau - t_\chi)^2$ in Equation (5.1), $\kappa(\tau + d - t_\chi)^2 - \nu$ would capture these traits. It is easy to see that, in equilibrium, the tip of the one-time customer is $\underline{t} = \{m + d - \frac{1}{2\kappa}\}^+$. Obviously, d must be low enough to ensure that the lowest tippers tip less than the average tip (m); $d < 1/(2\kappa)$. Here again, the social welfare losses can be minimized when $d \to 1/(2\kappa)$ as then, the tipping density must become concentrated at m. Similarly as for $\kappa \to +\infty$, there exists an arrival rate of repeat customers, Λ_r, that is strictly positive, but, small, such that one-time as well as all repeat customers join at approximately the optimal social joining rate, yielding arbitrarily small welfare losses.

[9] compared to any other non-anticipating regime in an M/M/1 with identical (effective) arrival rate, see Kingman (1962).

Finally, it follows immediately that ν simply increases the utility of the service in a model with tipping to $U + \nu$. For sufficient high ν, tipping (with a social norm) will yield a *higher* tipping wage than $\mu U(1 - \sqrt{c/(U\mu)})^2$. The latter is because the tipping process itself contributes to the welfare.

5.4 Concluding Remarks

Tips are an important component of a customer's cost of obtaining service, as well as an important component of the income of many low-paid service employees. In this chapter, we developed a model in which the service quality is service speed. As in the literature, we focus on one-time customers tipping behavior and show, that in the absence of *behavioral* motivations, these customers would never leave a tip after service. We introduce a social norm (like Azar (2004b)) in a queueing game with priorities (like Glazer and Hassin (1986)). To make insights sharp, *all* customers are homogeneous in terms of their service valuation (U) and waiting cost sensitivity (c). The only source of heterogeneity is the impact of the tip on waiting time; repeat customers (Λ_r arrival rate) can reduce their waiting time by tipping more, while one-time customers (Λ_o arrival rate) cannot.

While in the theoretical tipping literature the main focus is on the strategic interaction between the customer and service provider, in a capacitated environment with a social norm, as in our model, the interactions among customers are crucial. The service rate in our model is exogenously given (i.e. the service provider cannot increase the service rate with some effort), but, not unlike practice, the service provider can allocate priority to customers who are believed to tip the most. Our main performance measure is the tipping wage, that is, the rate at which the service provider earns a wage via tips (i.e. the joining rate times the average tip). The queuing literature, Edelson and Hildebrand (1975), Glazer and Hassin (1986), provides mechanisms to maximize the social welfare generated without tipping (e.g. setting a service fee). We compare the equilibrium tipping wage with the maximum social welfare. Our main findings, summarized in the Corollaries are the following:

- In the absence of a social norm, one-time customers do not tip. They only join when repeat customers join too, and the potential market of the latter is relatively small. Thus, as in the tipping literature, in our model, rational one-time customers are not motivated to tip in the absence of a social norm.
- In the absence of a social norm, customers over-join the service facility when the potential market of repeat customers is small, such that they all join, as well

as some one-time customers. Thus, contrary to the queueing literature with homogenous customers (in terms of preferences, U, c), the heterogeneity introduced by the repeat customers causes losses in the tipping wage. One-time customers, who determine the joining rate, do not internalize fully the waiting externalities their joining decision imposes on other customers. Repeat customers earn strictly positive rents that cannot be appropriated via tips.

- In the absence of a social norm, the tipping wage achieves the maximum social welfare when the potential market of repeat customers is sufficiently high such that not all of them join. Repeat customers join at the socially optimal rate, and earn zero rents, one-time customers don't join.

- In the presence of a social norm, one-time customers do *not* necessarily tip when they join. Contrary to the tipping literature (with an exogenous social norm), the existence of a social norm is *not* sufficient to motivate one-time customers to tip. Only when the service value is high enough and the potential market of repeat customers is neither too high nor too low, there exists an equilibrium in which one-time customers tip. When they tip, all repeat customers also join (and tip also).

- In the presence of a social norm, when the potential market for repeat customers is neither very small nor very large, at least two equilibria emerge; one in which one-time customers join and tip and another one in which they don't join. In the former equilibrium, the tipping wage is the highest, the demand is the lowest.

- In the presence of a strong social norm, the social welfare losses of the tipping wage are small when the potential market of repeat customers is small (as the latter earn rent) and one-time customers join and tip. A strong social norm makes the tipping distribution concentrated, which makes the social norm cost minimal.

We also find that when customers care about tipping *more* than other customers, equilibria exists in which social welfare losses are minimal as the latter reduces the spread in the tip. Finally, we find that when customers *enjoy* tipping, the tipping wage can exceed the maximum social welfare (without tipping).

The main driving forces in our model are (1) the tension between one-time and repeat customers and (2) the variability that competing repeat customers introduce. Factor (1) causes inefficiencies because when both types join, the repeat customers are able to earn strictly positive rent. The intuition is that these customers can influence better the service quality via their tip. The consequence is that the equilibrium joining rate is too high, compared with the socially optimal joining rate. Factor (2) causes inefficiencies

because the social norm is convex in the tip. Therefore, the social costs are always strictly positive. The intuition is that when (homogeneous) repeat customers tip for faster service, they try to differentiate themselves by tipping different amounts, generating variability in tips and therefore social cost.

According to Azar (2005a), if future service is a reason for tipping, the sensitivity of tips to service quality should be higher for repeating customers than for non-repeating ones. In our model, the sensitivity of tips to service quality is the same for repeat and one-time customers and yet, one-time customers might tip. According to Azar (2004b), to sustain a social norm on the long run, it is required that customers derive strictly positive benefits from tipping higher than others and from being perceived as generous when tipping. Otherwise, the tipping norm erodes over time. In our model, such benefit is not necessary; even when $\nu = 0$ and $d = 0$, the tipping norm does not erode (as there exist equilibria with one-time customers that tip). Our result is driven by the coexistence of one-time customers and repeat customers in a capacitated environment. Competition among the latter for fast service generates sufficient variation in the tips such that one-time customers feel compelled to tip, based on negative emotions only (i.e. tipping a different amount than others).

To the best of our knowledge, no prior research studies tipping for congested services. Adding a social norm to a queuing game is new, and so is adding a queue to a tipping game. While the main insights obtained are consistent with both queuing and tipping literatures, we believe new insights emerge: First of all, as most of the tipping literature with a social norm focuses on the employee-customer interaction, the demand is typically exogenously given (a single customer). In our model, demand is endogenous and our welfare analysis takes this into account. From the perspective of the tipped employee, it is not the average tip that matters, but, also the demand. Second, our social norm arises endogenously and is fundamentally driven by one parameter, κ, capturing the dis-utility of not conforming to the norm. Third, we characterize multiple equilibria with a social norm. This should be expected as norms are real actions (tips) based on beliefs (about other customers' tips). Multiple actions and beliefs might be consistent. In general, in queuing games as in Glazer and Hassin (1986), congestion externalities are negative (some customers avoid crowded facilties) resulting in an avoid-the-crowd based behavior of customers, and therefore in a unique equilibrium. Adding a social consumption stage, we observe positive externalities caused by mingling customers who tip lower amounts (one-time customers) with customers who tip higher amounts. These positive externalities may yield multiple equilibria. The existence of multiple equilibria might shed a different light on the evolution of social norms, see Azar (2004b). Fourth, the customer to customer interaction in the service consumption phase has significantly different implications for tipping than the customer to customer interaction in the social

consumption phase. In the service consumption phase, the finite capacity may trigger a tipping 'war' among customers, lifting up the average tip so much that some customers leave. The social consumption phase, on the other hand cannot provide such upward pressure on the tips, instead it reduces the dispersion of the tips. Finally, while in the tipping literature (with exogenous demand), the service quality is under-provided in equilibrium, we identify situations in which service quality (the average queue length) is over-provided (too short). The latter is because demand is endogenous; a tipping war among customers drives up the ex ante cost of joining, which may restrict the demand (joining rate), keeping the average queue length too low.

As with all models, ours is a simplification of reality, aimed to obtain insights based from first principles. To that end, we made simplifying assumptions. The following extensions of our model would be interesting, but, come at the cost of significant analytic complication: We assume that all customers are homogeneous in their delay sensitivity. As a consequence, a priority scheme can never generate more welfare than the FCFS under optimal price. Introducing heterogeneity in waiting costs might give a strict advantage to tipping when the social norm is weak. We also assumed that customers compare tips in the social consumption stage with one randomly selected other customer, without incorporating their actual service experience (as is done in Azar (2005a)). The social norm could depend on the comparison of the tip, conditional on the service experience (waiting time). The social norm component would then become more complicated. We leave these extensions for future research. In addition, our model did not capture a number of tipping behavioral motivations that play a role in practice. For example, we did not take into account altruistic reasons for tipping, or tipping as a means of approval by the server. We leave these for future research as well. Our modeling framework might be a rich one for further research, enhancing our theoretical understanding of tip behavior via formalized economic models. Better understanding will inform empirical work and, in the end, improve practice and impact public-policy issues of concern.

APPENDICES

Appendix A

Proofs for Chapter 2

A.1 Proof of Proposition 2.5

Proof. If $n_o = n_e$, it follows immediately by Proposition 2.4 that $n_m \leq n_o$. Suppose that $n_o \neq n_e$, then by Proposition 2.3, $n_o < n_e$. Since n_o is the maximum point of $\mathrm{E}\big(S^{(n)}\big)$,

$$\mathrm{E}\big(S^{(n_o)}\big) \geq \mathrm{E}\big(S^{(n)}\big), \quad \forall n \in [n_o, n_e]. \tag{A.1}$$

By assumption, $\mathrm{E}\big(S^{(n)} - M^{(n)}\big)$ is a non-decreasing sequence in $n \in [n_o, n_e]$, thus, we have

$$\mathrm{E}\big(S^{(n_o)}\big) - \mathrm{E}\big(M^{(n_o)}\big) \leq \mathrm{E}\big(S^{(n)}\big) - \mathrm{E}\big(M^{(n)}\big), \quad \forall n \in [n_o + 1, n_e]. \tag{A.2}$$

Subtracting (A.2) from (A.1) we arrive at

$$\mathrm{E}\big(M^{(n_o)}\big) \geq \mathrm{E}\big(M^{(n)}\big), \quad \forall n \in [n_o + 1, n_e],$$

which means that $n_m \notin [n_o + 1, n_e]$. From Assumption (A-i) and Proposition 2.4 it follows that $n_m \leq n_e$, and we conclude that $n_m \leq n_o$. □

A.2 Proof of Proposition 2.6

Proof. Recall that in such a system, the stationary number of customers at arrival instances (and therefore the offered position) is invariant to changes in service regime. Consider the system when customers follow threshold $n + 1$, but with the following modification: customers in the system are divided into two different (preemptive-resume) priority classes. Customers who are offered positions 1 to n join as regular priority customers, whereas customers offered position $n + 1$ join as second priority customers,

meaning they commence service only when there is no regular customer waiting in the system. Customers offered position $n+2$, from the $(n+1)$-threshold assumption, do not join at all. By construction, at any time point, there can be at most one second-priority customer in the system, thus, we tag this customer as the *standby* (or *transparent*) customer, similarly to Haviv (2013)§4.7.3. All regular customers are indifferent (in terms of sojourn time) to the existence of the standby customer.

It can be easily seen that with the above modification, the joining strategy of regular customers coincides with the n-threshold strategy, thus the number of regular customers in the system (at arrival instances) is given by $Q^{(n)}$. Since the prioritization scheme does not affect the stationary distribution of the process, the distribution of total number of customers in the system at arrivals, including the standby customer, is as of $Q^{(n+1)}$. Moreover, at any given moment, the number of awaiting standby customers is at most one, yielding $Q^{(n+1)} \in \{Q^{(n)}, Q^{(n)}+1\}$, and so, by Shaked and Shanthikumar (2007) Theorem 1.A.1, $Q^{(n)} \leq_{\mathrm{st}} Q^{(n+1)} \leq_{\mathrm{st}} Q^{(n)}+1$. □

A.3 Proof of Lemma 2.9

Proof. First, we note that from (2.4) and the definition of J_n and B_n,

$$\mathbf{1}_{J_{n+1}} = \mathbf{1}_{J_n} + \mathbf{1}_{B_n}\mathbf{1}_{A_n^c} = \mathbf{1}_{J_n}\mathbf{1}_{A_n} + \mathbf{1}_{A_n^c},$$

thus,

$$\begin{aligned}
\mathrm{E}\big(M^{(n+1)} - M^{(n)}\big) &= \mathrm{E}\big(u(n+1)\cdot\mathbf{1}_{J_{n+1}} - u(n)\cdot\mathbf{1}_{J_n}\big) \\
&= \mathrm{E}\left(u(n+1)\cdot\big(\mathbf{1}_{J_n} + \mathbf{1}_{B_n}\mathbf{1}_{A_n^c}\big) - u(n)\cdot\mathbf{1}_{J_n}\right) \\
&= u'(n)\cdot\Pr(J_n) + u(n+1)\cdot\Pr(B_n\cap A_n^c). \qquad \text{(A.3)}
\end{aligned}$$

In addition,

$$\begin{aligned}
\mathrm{E}\big(S^{(n+1)} - S^{(n)}\big) &= \mathrm{E}\left(u\big(Q^{(n+1)}\big)\cdot\mathbf{1}_{J_{n+1}} - u\big(Q^{(n)}\big)\cdot\mathbf{1}_{J_n}\right) \\
&= \mathrm{E}\left(\big(\mathbf{1}_{J_n}\mathbf{1}_{A_n} + \mathbf{1}_{A_n^c}\big)\cdot u\big(Q^{(n)}+\mathbf{1}_{A_n}\big) - \big(\mathbf{1}_{J_n}\mathbf{1}_{A_n} + \mathbf{1}_{J_n}\mathbf{1}_{A_n^c}\big)\cdot u\big(Q^{(n)}\big)\right) \\
&= \mathrm{E}\left(\mathbf{1}_{J_n}\mathbf{1}_{A_n}\left(u\big(Q^{(n)}+1\big) - u\big(Q^{(n)}\big)\right) + \mathbf{1}_{A_n^c}\left(u\big(Q^{(n)}\big) - \mathbf{1}_{J_n}u\big(Q^{(n)}\big)\right)\right) \\
&= \mathrm{E}\left(\mathbf{1}_{J_n}\mathbf{1}_{A_n}u'\big(Q^{(n)}\big)\right) + \mathrm{E}\left(\mathbf{1}_{B_n}\mathbf{1}_{A_n^c}u\big(Q^{(n)}\big)\right) \\
&= \mathrm{E}\left(\mathbf{1}_{J_n}\mathbf{1}_{A_n}u'\big(Q^{(n)}\big)\right) + u(n+1)\cdot\Pr(B_n\cap A_n^c). \qquad \text{(A.4)}
\end{aligned}$$

Subtracting (A.3) from (A.4), together with the definition of $D^{(n)}$ we achieve

$$\mathrm{E}\big(D^{(n+1)} - D^{(n)}\big) = \mathrm{E}\left(\mathbf{1}_{J_n}\mathbf{1}_{A_n}u'\big(Q^{(n)}\big)\right) - u'(n)\cdot\Pr(J_n),$$

which is equivalent to (2.5). □

A.4 Proof of Proposition 2.10

Proof. Suppose (2.6) holds. Note that since $u(k)$ is nonicreasing, $u'(k) \le 0$ for every k and therefore $\mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n\right) \le 0$ for every $n \in [n_o, n_e]$. Equation (2.6) can be then rewritten as

$$\mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n\right)\cdot\Pr(J_n \cap A_n) - u'(n)\cdot\Pr(J_n) \ge 0, \quad \forall n \in [n_o, n_e].$$

This, with (2.5), implies that

$$\mathrm{E}\big(D^{(n+1)} - D^{(n)}\big) \ge 0, \quad \forall n \in [n_o, n_e].$$

Thus, by Proposition 2.5, $n_m \le n_o$. Together with Proposition 2.3, the result is obtained. □

A.5 Proof of Corollary 2.11

Proof. Since $u(k)$ is convex decreasing, $u'(1) \le u'(k) \le 0$ for all $k \ge 1$. By the assumption,

$$\frac{u'(n)}{\Pr(A_n \mid J_n)} \le u'(1) \le \mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n\right), \quad \forall n \in [n_o, n_e]$$

where the second inequality follows since the right-hand side is a convex combination of $u'(k)$ such that $k \ge 1$. Thus (2.6) holds, and by Proposition 2.10, $n_m \le n_o \le n_e$. □

A.6 Proof of Corollary 2.12

Proof. Since $u(k)$ is decreasing and concave,

$$u'(n) \le u'(n-1) \le \cdots \le u'(1) \le 0.$$

Note that the random variable $\mathbf{1}_{A_n} u'\big(Q^{(n)}\big)$ takes values on the set $\{0\} \cup \{u'(k)\}_{k=1}^{n}$. It follows that for all n,

$$u'(n) \le \mathrm{E}\left(\mathbf{1}_{A_n} u'\big(Q^{(n)}\big) \;\middle|\; J_n\right) = \mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n\right) \cdot \Pr(A_n \mid J_n), \tag{A.5}$$

thus, since $\mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n\right) \le 0$, Equation (2.6) holds, and by Proposition 2.10, $n_m \le n_o \le n_e$. □

A.7 Proof of Lemma 2.13

Proof. By taking expectation on both sides of Equation (2.4) and rearranging we prove the first item. In addition,

$$\Pr(J_n \cap A_n) = \Pr(A_n) - \Pr(B_n \cap A_n) = \Pr(A_n) - \Pr(B_{n+1}),$$

and therefore,

$$\Pr(A_n \mid J_n) = \frac{\Pr(J_n \cap A_n)}{\Pr(J_n)} = \frac{\mathrm{E}\big(Q^{(n+1)}\big) - \mathrm{E}\big(Q^{(n)}\big) - \Pr(B_{n+1})}{\Pr(J_n)},$$

which proves the second item.

We note that

$$\Pr\big(Q^{(n)} = k\big) = \Pr\big(Q^{(n)} = k, A_n\big) + \Pr\big(Q^{(n)} = k, A_n^c\big), \quad k = 1, \ldots, n+1, \tag{A.6}$$

and that

$$\Pr\big(Q^{(n+1)} = k\big) = \Pr\big(Q^{(n)} = k-1, A_n\big) + \Pr\big(Q^{(n)} = k, A_n^c\big), \quad k = 2, \ldots, n+1. \tag{A.7}$$

Since $\{Q^{(n+1)} = 1\} = \{Q^{(n)} = 1, A_n^c\}$, we have, from (A.6) for $k = 1$,

$$\Pr\big(Q^{(n)} = 1, A_n\big) = \Pr\big(Q^{(n)} = 1\big) - \Pr\big(Q^{(n+1)} = 1\big).$$

Substituting (A.7) in (A.6) for $k = 1, \ldots, n+1$ we get

$$\Pr\big(Q^{(n)} = k, A_n\big) = \sum_{i=1}^{k} \Big(\Pr\big(Q^{(n)} = i\big) - \Pr\big(Q^{(n+1)} = i\big)\Big) = \Pr\big(Q^{(n)} \le k\big) - \Pr\big(Q^{(n+1)} \le k\big),$$

which completes the proof of the claim. □

A.8 Proof of Lemma 2.14

Proof. For every $k \leq n+1$

$$\begin{aligned}
\Pr\big(A_n \,\big|\, Q^{(n)} = k\big) &= \frac{\Pr\big(Q^{(n)} = k, A_n\big)}{\Pr\big(Q^{(n)} = k\big)} = \frac{\Pr\big(Q^{(n)} \leq k\big) - \Pr\big(Q^{(n+1)} \leq k\big)}{\Pr\big(Q^{(n)} = k\big)} \\
&= \frac{\Pr\big(Q^{(n+1)} > k\big) - \Pr\big(Q^{(n)} > k\big)}{\Pr\big(Q^{(n)} = k\big)} \\
&= \frac{\Pr(B_{n+1}) - \sum_{i=k+1}^{n+1} \big[\Pr\big(Q^{(n)} = i\big) - \Pr\big(Q^{(n+1)} = i\big)\big]}{\Pr\big(Q^{(n)} = k\big)} \\
&\leq \frac{\Pr(B_{n+1})}{\Pr\big(Q^{(n)} = k\big)} \leq \frac{\Pr(B_{n+1})}{\Pr(B_n)} = \frac{\Pr(B_n \cap A_n)}{\Pr(B_n)} = \Pr(A_n \mid B_n). \quad \text{(A.8)}
\end{aligned}$$

The second equality follows by the third item of Lemma 2.13. The first inequality follows since $\Pr\big(Q^{(n+1)} = k\big) \leq \Pr\big(Q^{(n)} = k\big)$, and the second inequality since $\Pr(B_n) \leq \Pr\big(Q^{(n)} = k\big)$, for all $k \leq n+1$. By the law of total probability and (A.8),

$$\begin{aligned}
\Pr(A_n) &= \Pr(A_n \mid J_n) \cdot \Pr(J_n) + \Pr(A_n \mid B_n) \cdot \Pr(B_n) \\
&= \sum_{k=1}^{n} \Pr\big(A_n \,\big|\, Q^{(n)} = k\big) \cdot \Pr\big(Q^{(n)} = k\big) + \Pr(A_n \mid B_n) \cdot \Pr(B_n) \leq \Pr(A_n \mid B_n).
\end{aligned}$$

We conclude that $\Pr(A_n) \leq \Pr(A_n \mid B_n)$, and therefore $\Pr(A_n \mid J_n) \leq \Pr(A_n)$. □

Appendix B

Proofs For Chapter 3

B.1 Proof of Lemma 3.1

Proof. We show that $\{\mathrm{E}\,(g(S_n))\}_{n=0}^{\infty}$ is convex by showing that the sequence of its differences is non-decreasing. Let $\tilde{X}$ be a random variable independent of all $\{X_i\}_{i=1}^{\infty}$ and identically distributed, then, for all $n \in \mathbb{N} \cup \{0\}$,

$$g(S_{n+1}) - g(S_n) \sim g(\tilde{X} + S_n) - g(S_n), \tag{B.1}$$

where the notation "$\sim$" stands for equality in distribution. Since g is convex and $S_{n-1} \leq S_n$, then for all $n \in \mathbb{N}$,

$$g(\tilde{X} + S_{n-1}) - g(S_{n-1}) \leq g(\tilde{X} + S_n) - g(S_n). \tag{B.2}$$

Taking expected values of both sides of (B.2) we arrive at

$$\begin{aligned}\mathrm{E}(g(S_n)) - \mathrm{E}(g(S_{n-1})) &= \mathrm{E}(g(\tilde{X} + S_{n-1})) - \mathrm{E}(g(S_{n-1})) \\ &\leq \mathrm{E}(g(\tilde{X} + S_n) - \mathrm{E}(g(S_n))) = \mathrm{E}(g(S_{n+1})) - \mathrm{E}(g(S_n)),\end{aligned}$$

concluding that $\{\mathrm{E}\,(g(S_n))\}_{n=1}^{\infty}$ is convex. □

B.2 Proof of Corollary 3.4

Proof. Recall $Q^{(n)}$, the offered position for an arriving customer when the threshold strategy is n. Define the following quantities

$$\beta_0 = 1; \quad \beta_k = \prod_{i=1}^{k} \frac{\lambda}{\mu + i\theta}, \quad k = 1, 2, \dots$$

Since the system is a standard birth-death process (M/M/1+M), using PASTA, it suffices to solve the steady-state equations to obtain the distribution of $Q^{(n)}$:

$$\Pr\big(Q^{(n)} = k\big) = \frac{\beta_{k-1}}{\sum_{j=0}^{n} \beta_j}, \quad k = 1, 2, \dots, n+1. \tag{B.3}$$

By Lemma 2.13 and (B.3), for any $k = 1, 2, \dots, n+1$,

$$\Pr\big(Q^{(n)} = k, A_n\big) = \sum_{i=1}^{k} \left(\frac{\beta_{i-1}}{\sum_{j=0}^{n} \beta_j} - \frac{\beta_{i-1}}{\sum_{j=0}^{n+1} \beta_j} \right) = \frac{\beta_{n+1} \sum_{i=0}^{k-1} \beta_i}{\left(\sum_{j=0}^{n} \beta_j\right) \cdot \left(\sum_{j=0}^{n+1} \beta_j\right)}, \tag{B.4}$$

and by (3.3), the definition of $u'(k)$ and the definition of β_k,

$$u'(k) = u(k+1) - u(k) = \frac{-R\mu\theta}{(\mu + k\theta)(\mu + (k+1)\theta)} = \frac{-R\mu\theta}{\lambda^2} \cdot \frac{\beta_{k+1}}{\beta_{k-1}}. \tag{B.5}$$

We shall show that

$$u'(n) \cdot \Pr(J_n) \le \mathrm{E}\left(u'\big(Q^{(n)}\big) \;\middle|\; J_n \cap A_n \right) \cdot \Pr(J_n \cap A_n), \tag{B.6}$$

and then, from Proposition 2.10 (recall that u' is nonpositive) we will derive the desired result. First note from (B.3) that

$$\Pr(J_n) = 1 - \Pr(B_n) = 1 - \frac{\beta_n}{\sum_{j=0}^{n} \beta_j}.$$

From (B.5) and (B.3) we have

$$u'(n) \cdot \Pr(J_n) = \frac{-R\mu\theta}{\lambda^2} \cdot \frac{\beta_{n+1}}{\beta_{n-1}} \cdot \frac{\sum_{j=0}^{n-1} \beta_j}{\sum_{j=0}^{n} \beta_j} = \frac{-R\mu\theta\beta_{n+1}}{\lambda^2 \sum_{j=0}^{n} \beta_j} \cdot \frac{1}{\Pr(B_{n-1})}.$$

Analyzing the right-hand side of (B.6) we get, by (B.4)

$$\begin{aligned}
&\mathrm{E}\left(u'\left(Q^{(n)}\right) \;\middle|\; J_n \cap A_n\right) \cdot \Pr\left(J_n \cap A_n\right) \\
&\quad = \sum_{k=1}^{n} u'(k) \cdot \Pr\left(Q^{(n)} = k, A_n\right) = \sum_{k=1}^{n} \frac{-R\mu\theta}{\lambda^2} \cdot \frac{\beta_{k+1}}{\beta_{k-1}} \cdot \frac{\beta_{n+1} \sum_{i=0}^{k-1} \beta_i}{\left(\sum_{j=0}^{n} \beta_j\right)\left(\sum_{j=0}^{n+1} \beta_j\right)} \\
&\quad = \frac{-R\mu\theta\beta_{n+1}}{\lambda^2 \sum_{j=0}^{n} \beta_j} \cdot \sum_{k=1}^{n} \frac{\beta_{k+1}}{\beta_{k-1}} \cdot \frac{\sum_{i=0}^{k-1} \beta_i}{\sum_{j=0}^{n+1} \beta_j} = \frac{-R\mu\theta\beta_{n+1}}{\lambda^2 \sum_{j=0}^{n} \beta_j} \cdot \sum_{k=1}^{n} \frac{\beta_{k+1}}{\sum_{j=0}^{n+1} \beta_j} \cdot \frac{1}{\Pr(B_{k-1})}.
\end{aligned}$$

Clearly,

$$\frac{-R\mu\theta\beta_{n+1}}{\lambda^2 \sum_{j=0}^{n} \beta_j} < 0,$$

therefore to show our aim (B.6) it suffices to show

$$\sum_{k=1}^{n} \frac{\beta_{k+1}}{\sum_{j=0}^{n+1} \beta_j} \cdot \frac{1}{\Pr(B_{k-1})} \leq \frac{1}{\Pr(B_{n-1})}.$$

Note that $\Pr(B_k)$ is positive and monotone decreasing in k, as it represents the blocking probability of a queue with threshold k. Thus, $\frac{1}{\Pr(B_k)}$ is monotone increasing. Now,

$$\sum_{k=1}^{n} \frac{\beta_{k+1}}{\sum_{j=0}^{n+1} \beta_j} \cdot \frac{1}{\Pr(B_{k-1})} \leq \sum_{k=1}^{n} \frac{\beta_{k+1}}{\sum_{j=1}^{n} \beta_{j+1}} \cdot \frac{1}{\Pr(B_{k-1})} \leq \frac{1}{\Pr(B_{n-1})},$$

where the first inequality evolves as we decrease the denominator and the second inequality follows as the second term is a convex combination of terms $\frac{1}{\Pr(B_{k-1})}$ such that $k \leq n$. □

B.3 Proof of Lemma 3.5

Proof. Let the state of the system be the position offered for an arriving customer. Since arrivals are Poisson we can, by the PASTA principle, analyze the state time-averaged distribution. To prove the first item, note that when $\lambda \leq \mu$ the sequence $\{\Pr\left(Q^{(n)} = k\right)\}_{k=1}^{n+1}$ is decreasing. This can be easily verified by writing down the balance equations, as done in Boudali and Economou (2012). However, we note that even in the extreme case when $\lambda = \mu$ and $\xi = 0$, the state distribution is uniform, and increasing either μ or ξ increases the drift towards states closer to state 1 (when the system is empty). Thus, $\Pr(B_n) \leq \Pr\left(Q^{(n)} = k\right)$ for all $k \leq n+1$.

For the second item, consider a pair of states $k, l \in \{1, \ldots, n+1\}$, and denote by $T_{k,l}$ the expected time elapsed from the moment of leaving state k until we first hit state l.

Then for every $k \in \{2, \ldots, n\}$,

$$\begin{aligned}\Pr\big(Q^{(n)} = k\big) &= \frac{\frac{1}{\lambda+\mu+\xi}}{\frac{1}{\lambda+\mu+\xi} + \frac{\lambda}{\lambda+\mu+\xi}T_{k+1,k} + \frac{\mu}{\lambda+\mu+\xi}T_{k-1,k} + \frac{\xi}{\lambda+\mu+\xi}T_{1,k}} \\ &= \frac{1}{1 + \lambda T_{k+1,k} + \mu T_{k-1,k} + \xi T_{1,k}}.\end{aligned}$$

The terms $T_{k-1,k}$ and $T_{1,k}$ do not depend on the threshold n, because the path from state l to state k when $l < k$ never crosses states greater than k. When in state $k+1$ the time to hit state k is either the time to the next catastrophe, or the time it takes to reduce the number of customers by one, the earliest among them, plus the hitting time from 1 to k if the catastrophe occurred first. Both the time to the next catastrophe and the hitting time from 1 to k are independent of the threshold n. The time it takes to reduce the system by one (given that a catastrophe did not occur) increases stochastically with n, for every $k \leq n$. Therefore $T_{k+1,k}$ is increasing with n and so, $\Pr\big(Q^{(n)} = k\big) \geq \Pr\big(Q^{(n+1)} = k\big)$. □

B.4 Proof of Lemma 3.6

Proof. Consider the case where $\mu = 0$, that is, service is infinitely long and customers only leave the system when a catastrophe occurs. By showing that $\Pr(A_n)$ is maximal for $\mu = 0$ we will obtain an upper bound for the term $\Pr(A_n)$ for any $\mu \geq 0$.

Note first that in the $\mu = 0$ case, the state transitions possible within one step are from state i to state $i+1$, $i \in \{1, \ldots, n\}$ and to state 1 (i.e., empty system). Writing down the balance equations for $\mu = 0$ we have

$$\begin{cases} \lambda \Pr\big(Q^{(n)} = 1\big) = \xi \Pr(B_n) + \xi \sum_{i=2}^{n} \Pr\big(Q^{(n)} = i\big), \\ (\lambda + \xi) \cdot \Pr\big(Q^{(n)} = k\big) = \lambda \Pr\big(Q^{(n)} = k-1\big), \quad k = 2, \ldots, n, \\ \xi \Pr(B_n) = \lambda \Pr\big(Q^{(n)} = n\big), \end{cases}$$

and

$$\Pr(B_n) + \sum_{i=1}^{n} \Pr\big(Q^{(n)} = i\big) = 1,$$

and their solution is easily found to be[1]

$$\Pr\big(Q^{(n)}=k\big)\Big|_{\mu=0}=\left(\frac{\lambda}{\lambda+\xi}\right)^{k-1}\cdot\left(\frac{\xi}{\lambda+\xi}\right),\quad k=1,\ldots,n,$$
$$\Pr(B_n)|_{\mu=0}=\left(\frac{\lambda}{\lambda+\xi}\right)^{n}.$$

In addition,

$$\Pr(A_n)|_{\mu=0}=\Pr(B_{n+1})|_{\mu=0}=\left(\frac{\lambda}{\lambda+\xi}\right)^{n+1},$$

because for $\mu=0$, the event A_n occurs only when the system is in state B_n and a customer arrives before the next catastrophe. Note that when $\mu=0$ and the system is empty, the standby customer (who is offered position $n+1$) enters the system if and only if $n+1$ arrivals (including her own) occurred before a catastrophe. However, when $\mu>0$ this is merely a necessary condition. Thus, the expected time between the exit of a standby customer from the system and the entrance of a new standby customer into the system is smaller when $\mu=0$ than in the case $\mu>0$. Moreover, when $\mu=0$, once the standby customer has entered the system, she only leaves the system by the next catastrophe, which in expectation will occur within $\frac{1}{\xi}$ units time. As for the $\mu>0$ case, the standby customer may leave when a catastrophe occurs or sooner due to service completion, thus, her expected sojourn time in the system when $\mu=0$ is longer than in the case $\mu>0$. We conclude therefore that

$$\Pr(A_n)\le\Pr(A_n)|_{\mu=0}=\left(\frac{\lambda}{\lambda+\xi}\right)^{n+1},$$

and, along with the assumption $\lambda\le\mu$,

$$\Pr(A_n)\le\left(\frac{\lambda}{\lambda+\xi}\right)^{n+1}\le\left(\frac{\mu}{\mu+\xi}\right)^{n+1}\le\left(\frac{\mu}{\mu+\xi}\right)^{n-1}.$$

□

B.5 Proof of Lemma 3.7

Proof. Consider two processes, $N^{(n)}(t)$ and $N^{(n+1)}(t)$, representing the number of customers in the system at time t, when operating under threshold n and $n+1$, respectively.

[1] To explain this, consider a customer who arrives at time t. Denote by t' the time of the last catastrophe that occurred before t ($t'<t$). This customer is offered position $k\in\{1,\ldots,n\}$ in the queue if *exactly* $k-1$ customers arrive during the time interval (t',t), and balks if *at least* n customers arrive. Since the process is time reversible, the length of the interval (t',t) is exponentially distributed with rate ξ. Thus, the number of customers arriving between (t',t) follows a *shifted* geometric distribution with parameter $\frac{\xi}{(\lambda+\xi)}$, hence, the distribution of $Q^{(n)}$ (By shifted geometric distribution we refer to the distribution of the number *failures* before the first success in a sequence of Bernoulli trials).

Since the arrival process is Poisson, we have by the PASTA property that $Q^{(n)} = N^{(n)}+1$, where $N^{(n)}$ is the limiting distribution of $N^{(n)}(t)$ when $t \to \infty$. It suffices therefore to show that $N^{(n)} \leq_{\text{st}} N^{(n+1)} \leq_{\text{st}} N^{(n)} + 1$.

First, we show that $N^{(n)} \leq_{\text{st}} N^{(n+1)}$. We construct a coupling as follows: We assume that at time $t = 0$, $N^{(n)}(0) = N^{(n+1)}(0) = 0$, and the arrival process to both systems is identical. Each customer arrives at both systems simultaneously, retaining her own service demand which is drawn independently from some general distribution. Suppose a customer arrives at time t_0, than we assume she takes her action according to the following set of implications:

- If $N^{(n)}(t_0) < n$ and $N^{(n+1)}(t_0) < n + 1$, she joins both systems.
- If $N^{(n)}(t_0) = n$ and $N^{(n+1)}(t_0) = n + 1$, she balks from both systems.
- If $N^{(n)}(t_0) = n$ and $N^{(n+1)}(t_0) < n + 1$, she balks from the n-systems, and joins the $(n + 1)$-system.
- If $N^{(n)}(t_0) < n$ and $N^{(n+1)}(t_0) = n + 1$, this means that among the n *waiting* customers (those not being served) in the $(n+1)$-system, there exists at least one that at time t_0 is not present in the n-system (if there are more than one such customers, we pick the oldest in the system). This customer leaves the $(n + 1)$-system immediately, and the newly arriving customer joins at the back of the queue at both systems. This assumption is w.l.o.g, because the service demands of all customers are i.i.d.

With this construction, it can be observed that whenever a customer joins the n-system, she also joins the $(n+1)$-system. Moreover, the relative ordering of the customers joining the n-system is the same as their ordering at the $(n+1)$-system. Thus, if a customer at time t is present in the n-system, she is also present in the $(n + 1)$-system. Therefore, $N^{(n)}(t) \leq N^{(n+1)}(t)$ for all $t \in [0, \infty)$, implying that $N^{(n)} \leq_{\text{st}} N^{(n+1)}$.

We turn now to proving $N^{(n+1)} \leq_{\text{st}} N^{(n)} + 1$. Recall that when arrivals are Poisson, the number in the system embedded at time instants just after service completions forms a Markov Chain. When the threshold is n, the state space of this chain is the set $\{0, 1, \ldots, n - 1\}$. Denote by $B_i^{(n)}$ the state of the chain a moment after the i-th service completion, when the threshold is n. Let X_i be the (random) number of arrivals during the i-th service period. Then,

$$B_{i+1}^{(n)} | \{B_i^{(n)} = 0\} = \min\{X_{i+1}, n - 1\}, \tag{B.7}$$

and for $k \in \{1, \ldots, n-1\}$

$$B_{i+1}^{(n)} | \{B_i^{(n)} = k\} = \min\{k - 1 + X_{i+1}, n - 1\}. \tag{B.8}$$

Assume that the processes $\{B_i^{(n)}\}_{i=1}^{\infty}$ and $\{B_i^{(n+1)}\}_{i=1}^{\infty}$ are both induced by the same sequence $\{X_i\}_{i=1}^{\infty}$, and that at stage 1, $B_1^{(n)} = B_1^{(n+1)} = 0$. Consider some stage i,

- If $B_i^{(n)} = 0$ and $B_i^{(n+1)} \in \{0, 1\}$, then by (B.7) and (B.8) (with $k = 1$)

$$B_{i+1}^{(n)} = \min\{X_{i+1}, n-1\} \leq \min\{X_{i+1}, n\} = B_{i+1}^{(n+1)} \leq \min\{X_{i+1}, n-1\}+1 = B_{i+1}^{(n)}+1.$$

- If $B_i^{(n)} = k > 0$ and $k \leq B_i^{(n+1)} \in \{k, k+1\}$ then by (B.8),

$$\begin{aligned} B_{i+1}^{(n)} &= \min\{k - 1 + X_{i+1}, n - 1\} \leq \min\{k - 1 + X_{i+1}, n\} \leq B_{i+1}^{(n+1)} \\ &\leq \min\{k + X_{i+1}, n\} = \min\{k - 1 + X_{i+1}, n - 1\} + 1 = B_{i+1}^{(n)} + 1. \end{aligned}$$

By induction we deduce for all $i \in \mathbb{N}$, $B_i^{(n)} \leq B_i^{(n+1)} \leq B_i^{(n)} + 1$, therefore

$$B^{(n)} \leq_{\text{st}} B^{(n+1)} \leq_{\text{st}} B^{(n)} + 1, \tag{B.9}$$

where $B^{(n)} = \lim_{i\to\infty} B_i^{(n)}$. Denoting $\beta_k^{(n)} = \Pr\big(B^{(n)} = k\big)$, an equivalent expression for (B.9) is given by

$$\sum_{j=0}^{k} \beta_j^{(n)} \geq \sum_{j=0}^{k} \beta_j^{(n+1)} \geq \sum_{j=0}^{k-1} \beta_j^{(n)}, \quad k = 0, 1, \ldots, n. \tag{B.10}$$

(for $k = 0$ we define the sum in the right-hand-side of the inequality by 0). In particular, setting $k = 0$ in (B.10) implies $\beta_0^{(n)} \geq \beta_0^{(n+1)}$, and therefore, with $\bar{x}$ being the mean service time and $\rho = \lambda\bar{x}$,

$$0 \leq \frac{1}{\beta_0^{(n)} + \rho} \leq \frac{1}{\beta_0^{(n+1)} + \rho}. \tag{B.11}$$

Equations (B.10) and (B.11) yield

$$\frac{1}{\beta_0^{(n+1)} + \rho} \left(\sum_{j=0}^{k} \beta_j^{(n+1)} \right) \geq \frac{1}{\beta_0^{(n)} + \rho} \left(\sum_{j=0}^{k-1} \beta_j^{(n)} \right), \quad k = 0, 1, \ldots, n. \tag{B.12}$$

By Cohen (1982)§3.6.3, we have for all $j \in \{0, n-1\}$

$$\Pr\big(N^{(n)} = j\big) = \frac{\beta_j^{(n)}}{\beta_0^{(n)} + \rho},$$

thus, by (B.12), we obtain $\Pr\big(N^{(n+1)} \leq k\big) \leq \Pr\big(N^{(n)} \leq k-1\big)$ for all $k \in \{0, 1, \dots n\}$, therefore $N^{(n+1)} \leq_{\text{st}} N^{(n)} + 1$.

Overall, we proved that $N^{(n)} \leq_{\text{st}} N^{(n+1)} \leq_{\text{st}} N^{(n)} + 1$, and as $Q^{(n)} = N^{(n)} + 1$, we conclude that $Q^{(n)} \leq_{\text{st}} Q^{(n+1)} \leq_{\text{st}} Q^{(n)} + 1$, satisfying Assumption (A-ii). □

Appendix C

Proofs for Chapter 4

C.1 Proof of Proposition 4.1

Proof. A necessary and sufficient condition for stability is $\hat{\rho}(p,\rho) < 1$, i.e.,

$$p > \frac{\rho - 1}{\rho(2-\rho)} = \underline{p}\,.$$

To Find the maximum throughput we study the case $p = 1$, which means everyone senses S_L before joining the queue. Then, from (4.7), the stability criterion for $p = 1$ gives:

$$\hat{\rho}(1,\rho) = \frac{\rho^2}{1+\rho} < 1 \quad \Leftrightarrow \quad \rho^2 - \rho - 1 < 0 \quad \Leftrightarrow \quad \rho < \frac{1+\sqrt{5}}{2} = \varphi$$

Thus, for $\rho \geq \varphi$ and $p = 1$ the system is not stable. Note that $\hat{\rho}(p,\rho)$ is monotone increasing in p. Consequently, if $\rho \geq \varphi$ then the system is not stable for every $p \in [0,1]$, otherwise it is sable if and only if $p > \underline{p}$. □

C.2 Proof of Proposition 4.2

Proof. Proving the continuity of $\mathrm{E}[X \mid Y = 0]$ is immediate as, referring to (4.4), (4.5) and (4.6),

$$\mathrm{E}[X \mid Y = 0] = \frac{1}{\Pr(Y=0)} \sum_{i=0}^{\infty} iP_{i,0} = (1+p\rho) \sum_{i=0}^{\infty} iP_{i,0}\,,$$

which is a countable sum of continuous functions in $p \in [0,1]$.

In order to prove monotonicity, we examine the transitions between states in two coupled systems, $\Omega := \{S_Q, S_L\}$ and $\Omega' := \{S'_Q, S'_L\}$, under the same sequence of events. Systems

Ω and Ω' are identical except that in Ω the sensing probability is p, as opposed to Ω', where the sensing probability is p', and $p < p'$.

Denote by $(X(t), Y(t))$ and $(X'(t), Y'(t))$ the states at time t of systems Ω and Ω' respectively. Since for all $p \in [0, 1]$ the state $(0, 0)$ in the Markov chain is recurrent, we can assume w.l.o.g. that at time $t = 0$,

$$(X(0), Y(0)) = (X'(0), Y'(0)) = (0, 0) .$$

Showing that

$$\forall t \in [0, \infty) : \mathrm{E}[X'(t) \mid Y'(t) = 0] \leq \mathrm{E}[X(t) \mid Y(t) = 0] \tag{C.1}$$

will complete the proof.

We begin by attaching three variables to each customer: Let $\{(T_i, \tau_i, u_i)\}_{i \in \mathbb{N}}$ represent the *arrival time*, *service duration* and *sensing aversion* of the i-th customer respectively, and for all $i \in \mathbb{N}$:

$$T_{i+1} - T_i \sim \mathrm{Exp}(\Lambda) , \quad \tau_i \sim \mathrm{Exp}(\mu) , \quad u_i \sim \mathrm{Uniform}[0, 1] . \tag{C.2}$$

Each customer arrives to both systems simultaneously. Customer i arriving at Ω senses if and only if $u_i \leq p$ while arriving at Ω' he/she senses if and only if $u_i \leq p'$. This ensures that the sensing probability is p in system Ω and p' in system Ω', and as a matter of fact, implies that customers who sense in Ω also sense in Ω'.

We assign each customer one of three types based on the values of their sensing aversion: Customer i will be assigned

- *type-L* (stands for "Low") if $u_i \in [0, p]$,
- *type-M* (stands for "Medium") if $u_i \in (p, p']$,
- *type-H* (stands for "High") if $u_i \in (p', 1]$.

By definition, upon arrival, an H-type customer joins both S_Q and S'_Q, the shared queues in Ω and Ω' respectively, an M-type customer joins S_Q but senses S'_L, whereas an L-type customer senses both S_L and S'_L.

W.l.o.g., we modify the service regime as follows:

(i) Arriving at Ω (Ω') when S_L (S'_L) is busy, customer i preempts the customer in service in S_L (S'_L) with no regard to type, and the preempted customer is routed

to continue the service in S_Q (S'_Q). Arriving at Ω (Ω') when S_L (S'_L) is empty, customer-i's actions follow by type.

(ii) The shared queue, subsystem S_Q (S'_Q), is a preemptive resume LCFS-PR queue, meaning that customers are served following a preemptive last-come-first-served discipline, and the preempted service resumes from the last point it stopped.

Lemma C.1. *Assumptions (i) and (ii) do not affect the stationary probabilities of the systems.*

Proof. We shall show that Lemma C.1 holds for system Ω and the proof for Ω' is identical.

Let i and $i+1$ be a successive pair of customers and suppose that customer $i+1$ preempts customer i in subsystem S_L. Denote $\Delta_{i,i+1} := T_{i+1} - T_i$ the time difference between their arrivals. Then, as a result of the memoryless property of the exponential distribution, the service duration of customer $i+1$ (i.e., τ_{i+1}) and the residual service of customer i (i.e., $\tau_i - \Delta_{i,i+1} \mid \tau_i > \Delta_{i,i+1}$) are independent and identically distributed. This incident of preemption is equivalent to the joining of the new comer to subsystem S_Q, the shared queue, when S_L is busy. Hence, assumption (i) does not influence the transition probabilities and the stationary probabilities remain as before.

The validity of assumption (ii) can be explained by the fact that in a work-conserving preemptive resume system with exponential service duration, as sustains in S_Q, the queue length is independent of the service regime. Consequently, assumptions (i) and (ii) do not restrict the generality of the model. □

Exploiting assumptions (i) and (ii) we immediately achieve several properties that will be later used in the proof:

Lemma C.2. *Under assumptions (i) and (ii), the following arguments hold:*

(a) *At any moment, if S_L (S'_L) is busy, then the last customer to arrive until that moment is in S_L (S'_L).*

(b) *Both in Ω and in Ω', all customers begin service at the moment of arrival, whether they sense or not.*

(c) *Customers beginning service in S_L upon arrival, begin service in S'_L as well.*

(d) *Customers beginning service in S'_Q upon arrival, begin service in S_Q as well.*

(e) *The sojourn time of customer i in S'_Q is no longer than the sojourn time of i in S_Q.*

Proof.

(a) Suppose customer i, which arrived at T_i, is the last customer until time t, and i is not in S_L at time t. From assumption (i), there are only two possibilities: either S_L was idle at time T_i and i did not sense it (i.e., i is of type H or M), or i joins S_L and completes service before time t. Hence, in both scenarios, S_L stays idle at time t. The proof for Ω' is similar.

(b) In system Ω, customers at their moment of arrival either begin service immediately in S_L, or, by assumption (ii), join the head of the shared queue, thus beginning service in S_Q. The proof for Ω' is similar.

(c) Recall that at time $t = 0$ both S_L and S'_L are empty. Only upon the arrival of an L-type customer will the state of S_L alter. Arriving at Ω', an L-type customer joins S'_L, regardless of the server's state. Suppose the last customer has joined both S_L and S'_L. By assumption (i), this tagged customer can either complete service in S_L and S'_L simultaneously (so S_L and S'_L become empty) or be preempted in the two subsystems by the next customer to come. As for the latter case, it holds that the preempting customer joins both S_L and S'_L. Thus, it follows by induction that customers who join S_L also join S'_L.

(d) This property is immediately achieved as the *modus tollens* form of property (c).

(e) It results from assumption (ii) that the sojourn time of each customer in the shared queue depends only on future arrivals. Combining this with property (d) we deduce (e).

□

With property (c) and the simultaneousness of occurrences we get:

$$\forall t \in [0, \infty) : \{Y(t) = 1\} \Rightarrow \{Y'(t) = 1\}\,, \text{ and } \{Y'(t) = 0\} \Rightarrow \{Y(t) = 0\}\,. \tag{C.3}$$

Moreover, from properties (d) and (e) we see that, if at an arbitrary point in time there is a customer waiting in S'_Q, then the same customer will also be waiting in S_Q, not leaving Ω before leaving Ω'. Therefore,

$$\forall t \in [0, \infty) : X'(t) \leq X(t)\,, \tag{C.4}$$

implying that the process representing the number of customers in S_Q over time stochastically dominates the one representing the number of customers in S'_Q.

For the sake of simplifying the notation, when there is no ambiguity we denote $(X(t), Y(t))$ and $(X'(t), Y'(t))$ by (X, Y) and by (X', Y') respectively. On utilizing (C.4), we get

$$\mathrm{E}[X' \mid Y' = 0] \leq \mathrm{E}[X \mid Y' = 0] \,.$$

To prove our claim (C.1) it suffices to prove:

$$\mathrm{E}[X \mid Y' = 0] \leq \mathrm{E}[X \mid Y = 0] \,. \tag{C.5}$$

Denote by i the last customer who arrived by time t, and let

$$A(t) := \{i \text{ joined } S_Q \text{ and } T_i + \tau_i > t\}$$

express the event "i joined the queue in Ω and was still in the system by time t". For i being the last customer, by property (b) we get that $A(t)$ occurred only if i was still in service in S_Q by time t. Therefore from property (a),

$$\forall t \in [0, \infty) : A(t) \Rightarrow \{Y(t) = 0\} \,. \tag{C.6}$$

Recall that by property (a) if there is a customer in S'_L at time t then it must be i (the last customer arrived). So, assuming that $Y(t) = 0$ and $Y'(t) = 1$, we have that customer i, arriving at Ω, joined S_Q, and at the same time, arriving at Ω', began service in S'_L. As i was the last customer until time t, knowing that $Y'(t) = 1$ implies that $T_i + \tau_i > t$. Hence,

$$\forall t \in [0, \infty) : \{\{Y'(t) = 1\} \cap \{Y(t) = 0\}\} \Rightarrow A(t) \,. \tag{C.7}$$

Lemma C.3. *Conditioned on $A(t)$, $X(t)$ and $Y'(t)$ are independent random variables.*

Proof. Let t be an arbitrary point in time and i be the last customer to arrive by that time. Assuming that $A(t)$ had happen, it means that i was still in Ω by time t. The residual service of customer i at time t is an exponential variable with parameter μ. Thus, we can assume w.l.o.g. that i's arrival (at time T_i) occurred a short moment before t. $A(t)$ can be rewritten then as the event when a new busy period in S_Q began right before t (i.e., $t = T_i^+ := T_i + \varepsilon$, where ε is a small positive infinitesimal quantity).

Denote T_n^+ the moment after customer n's arrival. To prove Lemma C.3 we shall show by induction that $\forall n \in \mathbb{N}$, $X(T_n^+)$ and $Y'(T_n^+)$ are conditionally independent given $A(T_n^+)$.

Presume that $A(T_n^+)$ is satisfied. Having that n joined S_Q means that n is not of type L, and, by (C.6), $Y(T_n^+) = 0$. Define $T_n^- := T_n - \varepsilon$, where ε is a small positive infinitesimal

quantity, the moment before n's arrival. Then,

$$X(T_n^+)|A(T_n^+) = X(T_n^-) + 1\,. \tag{C.8}$$

Regarding n's arrival at Ω', there are two possibilities:

- Upon n's arrival, S'_L was empty (i.e., $Y'(T_n^-) = 0$).
- Upon n's arrival, S'_L was busy (i.e., $Y'(T_n^-) = 1$).

To the extent that $A(T_n^+)$ holds and $Y'(T_n^-) = 0$, n's action (and therefore the state of S'_L) is determined one to one by n's type (M or H):

$$Y'(T_n^+) \mid A(T_n^+) = \begin{cases} 0, & \text{iff } n \text{ is of type-H,} \\ 1, & \text{iff } n \text{ is of type-M;} \end{cases} = \begin{cases} 0 \text{ w.p. } \frac{1-p'}{1-p}, \\ 1 \text{ w.p. } \frac{p'-p}{1-p}. \end{cases}$$

Since i's type does not depend on previous events in the system it follows from (C.8) that $X(T_n^+)$ and $Y'(T_n^+)$ are independent conditioned on $A(T_n^+)$.

To the extent that $A(T_n^+)$ holds and $Y'(T_n^-) = 1$, with assumption (i), n must have preempted his/her predecessor, $n-1$, in S'_L, simultaneously as joining S_Q. This is possible only if customer $n-1$ remained in service until customer n's arrival, and indicates that

(a) Customer $n-1$ joined S_Q upon arriving at Ω, and $Y(T_{n-1}^+) = 0$.

(b) Customer $n-1$ joined S'_L upon arriving at Ω', and $Y'(T_n^+)|A(T_n^+) = Y'(T_{n-1}^+) = 1$.

(c) Customer n had arrived before $n-1$ left S_Q, and $X(T_n^+)|A(T_n^+) = X(T_{n-1}^+) + 1$.

From (a), (b) and (C.7) we deduce that $A(T_{n-1}^+)$ holds. Followed by the induction hypothesis, $X(T_{n-1}^+)$ and $Y'(T_{n-1}^+)$ are independent conditioned on $A(T_{n-1}^+)$. Thus, from (b) and (c) we conclude that $X(T_n^+)$ and $Y'(T_n^+)$ we are also independent conditioned on $A(T_{n-1}^+)$.

It is left to show that $X(T_1^+)$ and $Y'(T_1^+)$ are conditionally independent given $A(T_1^+)$. We can assume, w.l.o.g., that $Y'(T_1) = 0$ (otherwise we observe the last customer that arrived before T_1). Thus, as discussed, cusotmer 1's action is equivalent to type, which is a random variable independent of $X(T_1^+)$. □

Define the operator $\mathrm{E}_{Y=0}[\bullet] := \mathrm{E}[\bullet \mid Y = 0]$. Combining Lemma C.3 with (C.7) we arrive at

$$\mathrm{E}_{Y=0}[X \mid A] = \mathrm{E}_{Y=0}[X \mid A, Y' = 1] = \mathrm{E}_{Y=0}[X \mid Y' = 1]\,. \tag{C.9}$$

Since $A(t)$ indicates the beginning of a new busy period at time t

$$\mathrm{E}_{Y=0}[X] \leq \mathrm{E}_{Y=0}[X \mid A]\,.$$

Upon substituting (C.9) we get

$$\mathrm{E}_{Y=0}[X] \leq \mathrm{E}_{Y=0}[X \mid Y' = 1]\,. \tag{C.10}$$

By the law of total expectation we have

$$\begin{aligned} \mathrm{E}_{Y=0}[X] \; &= \mathrm{E}_{Y=0}[X \mid Y' = 0] \cdot \Pr(Y' = 0 \mid Y = 0) + \\ &\quad\; \mathrm{E}_{Y=0}[X \mid Y' = 1] \cdot \Pr(Y' = 1 \mid Y = 0) \\ &= \mathrm{E}[X \mid Y' = 0] \cdot \Pr(Y' = 0 \mid Y = 0) + \\ &\quad\; \mathrm{E}_{Y=0}[X \mid Y' = 1] \cdot \Pr(Y' = 1 \mid Y = 0)\,, \end{aligned} \tag{C.11}$$

where the second equality follows from (C.3). Note that the right-hand side of (C.11) is a convex combination of $\mathrm{E}[X \mid Y' = 0]$ and $\mathrm{E}_{Y=0}[X \mid Y' = 1]$. Amalgamating (C.10) and (C.11) we attain

$$\mathrm{E}[X \mid Y' = 0] \leq \mathrm{E}_{Y=0}[X]\,,$$

thus confirming (C.5). This completes the proof of Proposition 4.2. □

C.3 Proof of Proposition 4.3

Proof. $p \in [0, 1]$ is an equilibrium strategy if no individual can benefit from choosing any alternative strategy. Recall C_S and C_N as defined in (4.8). Note that:

$$\mathrm{E}\left[X\right] = \Pr(Y = 1) \cdot \mathrm{E}\left[X \mid Y = 1\right] + \Pr(Y = 0) \cdot \mathrm{E}\left[X \mid Y = 0\right]\,,$$

and therefore,

$$\frac{1}{c_s}(C_N(p) - C_S(p)) = \gamma \Pr(Y = 0) \cdot \mathrm{E}[X|Y = 0] - 1. \tag{C.12}$$

Clearly, $\Pr(Y = 0)$ is positive monotone decreasing in p and by Proposition 4.2, $\mathrm{E}[X|Y = 0] = 1$ is positive non-increasing in p, thus, $C_N(p) - C_S(p)$ is decreasing in p. We have then that the conditions for equilibrium are as follows:

- If $C_N(0) \leq C_S(0)$ then $p = 0$ is an equilibrium strategy.
- If $C_N(1) \geq C_S(1)$ then $p = 1$ is an equilibrium strategy.
- If for some $p \in (0,1)$ $C_N(p) = C_S(p)$ then p is an equilibrium strategy.

Other cases are infeasible considering that $C_N(p)$ and $C_S(p)$ are continuous and $C_N(p) - C_S(p)$ is a monotone function.

Suppose first that there exists a value of $p \in (0,1)$ such that $C_N(p) = C_S(p)$.

From (C.12) and (4.4), $p_e = p \in (0,1)$ if and only if:

$$\gamma \mathrm{E}\left[X \mid Y = 0\right] = 1 + p\rho\,. \tag{C.13}$$

Note that the function $1 + p\rho$ is continuous and strictly monotone increasing in p, and therefore, using Proposition 4.2, we deduce that if there exists a solution p_e to (C.13) then it is unique, implying that p_e is a unique equilibrium strategy. We shall mention in passing that since $C_N(p) - C_S(p)$ is decreasing, then the existence of $p_e \in (0,1)$ implies that $C_N(0) > C_S(0)$ and that $C_N(1) < C_S(1)$.

Suppose now that there is no $p \in (0,1)$ satisfying (C.13)and that $C_N(1) \geq C_S(1)$. Equivalently, for all $p \in [0,1)$, $C_N(p) > C_S(p)$. This means that the expected cost for an individual is smaller when sensing then when not sensing, regardless of other customers' strategy. Thus, sensing is a dominant strategy for all individuals, and therefore $p_e = 1$ is a unique equilibrium strategy. Similarly, $p_e = 0$ is the unique equilibrium strategy if and only if $C_N(0) \leq C_S(0)$. □

C.4 Proof of Proposition 4.4

Proof. Assuming that $\hat{\rho}(0,\rho) = \rho < 1$, from (C.13), $p = 0$ is an equilibrium strategy if and only if:

$$\gamma \mathrm{E}[X \mid Y = 0] \leq 1\,.$$

Note that in this case the queue is an M/M/1 queue, and

$$\mathrm{E}[X \mid Y = 0] = \mathrm{E}[X] = \frac{\rho}{1-\rho}\,.$$

Thus, a necessary and sufficient condition for $p_e = 0$ is:

$$\gamma \frac{\rho}{1-\rho} \leq 1 \quad \Leftrightarrow \quad \rho \leq \frac{1}{1+\gamma}\,.$$

□

C.5 Proof of Proposition 4.5

Proof. For $p = 1$, $g(z)$ is a polynomial of the second degree, and

$$z_0 = \frac{1}{\rho}\left(1 - \frac{1}{\sqrt{\rho+1}}\right) = \frac{1}{(1+\theta)\theta}\,.$$

With (4.10) we obtain

$$P_{0,0} = \frac{1+\theta-\theta^2}{\theta}\,.$$

Substituting this result in (4.12), after basic algebraic manipulations we get

$$\mathrm{E}\left[X \mid Y=0\right] = \frac{(\theta-1)^2(\theta+1)}{1+\theta-\theta^2} = \frac{1-\theta-\theta^2+\theta^3}{1+\theta-\theta^2}\,.$$

Recall that, by (4.4), $p_e = 1$ if and only if $\gamma \mathrm{E}\left[X \mid Y=0\right] \geq 1+\rho$. We have then that $p_e = 1$ if and only if

$$\gamma\left(\frac{(\theta-1)^2(\theta+1)}{1+\theta-\theta^2}\right) \geq \theta^2\,. \tag{C.14}$$

Note for every $\rho \in (0, \varphi)$, both $\theta > 0$ and $1+\theta-\theta^2 > 0$, thus, we can multiply each sides of the inequality in (C.14) by $1+\theta-\theta^2$ and divide by $(\theta-1)^2(\theta+1)$ to get the desired result. □

C.6 Proof of Proposition 4.6

Proof. Let γ and ρ be the fixes parameters of the system. If $p_e \in (0,1)$ is the equilibrium strategy, it must hold that

$$C_S(p_e) = C_N(p_e) \quad \Leftrightarrow \quad \gamma \mathrm{Pr}(Y=0) \cdot \mathrm{E}[X \mid Y=0] = 1, \tag{C.15}$$

where the terms $\mathrm{Pr}(Y=0)$ and $\mathrm{E}[X \mid Y=0]$ are calculated from ρ and the sensing probability. By (4.4) and Proposition 4.2, both $\mathrm{Pr}(Y=0)$ and $\mathrm{E}[X \mid Y=0]$ are positive and non-increasing in the sensing probability, thus, so is their product. So, if γ is increased, then in order to satisfy the equality in (C.15), one has to reduce $\mathrm{Pr}(Y=0) \cdot \mathrm{E}[X \mid Y=0]$ by increasing the sensing probability. □

C.7 Proof of Proposition 4.7

Proof. To prove that $F(p)$ is convex we shall show that its derivative is monotone non-decreasing in p. Given that $c_s = 0$ and $\frac{c_w}{\mu} = 1$, the expected cost of a customer ($F(p)$) is the expected number of customers ahead of him/her, which varies depending on his/her action (sensing or not sensing). Let $\delta > 0$ such that $p + \delta \leq 1$, and define $D(p)$ as the expected difference between the cost for a customer when the sensing probability is $p+\delta$ and when the sensing probability is p. Namely,

$$D(p) = F(p + \delta) - F(p) \,. \tag{C.16}$$

It is clear that the more customers sense, the greater the system utilization is, and since $c_s = 0$, the expected cost for a customer (equivalently, the social cost) decreases. Thus, for all $p \in [0, 1]$, $D(p) \leq 0$. Note that $\lim_{\delta \to 0} \frac{D(p)}{\delta} = (\frac{d}{dp})F(p)$, which can be interpreted as the marginal social cost of a sensing customer when the sensing probability is p, normalized in the arrival rate, λ.

Lemma C.4. *$D(p)$ is monotone non-decreasing in p.*

Proof. As in the proof of Proposition 4.2, recall $\Omega = \{S_Q, S_L\}$ and $\Omega' = \{S'_Q, S'_L\}$ with the coupled states $(X(t), Y(t))$ and $(X'(t), Y'(t))$ and sensing probabilities p and p' $(p < p' < 1)$ respectively. We use $\{(T_i, \tau_i, u_i)\}_{i \in \mathbb{N}}$ - the set of variables defining the queueing process as in (C.2), with T_i the arrival time, τ_i the service duration and u_i the sensing aversion of customer i. We also assume (i) and (ii) as in Lemma C.1. Our purpose is to show $D(p) \leq D(p')$, that is, the marginal cost of increasing p by an amount δ, is not greater then the marginal cost of increasing p' by δ.

In the next step of the proof, we point out how increasing the number of sensing customers influences the expected cost of customers within the two coupled systems, Ω and Ω'. To this end, in each of the two systems we tag a customer that initially joins the shared queue and analyze how the social cost changes when he/she senses. Let i be a customer such that $u_i > p'$ (such customer exists because $p' < 1$). This means that from the outset i did not sense, neither in Ω nor in Ω'. By changing i's action into sensing we assure that the population of sensing customers, both in Ω and in Ω' strictly increases. Thus, to measure the impact in the two systems of a customer switching from not sensing to sensing, we can, w.l.o.g., assume that it is customer i. The fact that each customer that senses in Ω also senses in Ω' is still valid, regardless of whether i senses in both systems or i joins the shared queue in both systems. This ensures that Lemma C.2 holds in each of the two cases, i senses and i does not sense, and our aim is to investigate systems Ω and Ω', subject to each one of these cases. Define:

- D_i - The social cost difference in system Ω between the case i senses and the case he/she does not sense.
- D'_i - The social cost difference in system Ω' between the case i senses and the case he/she does not sense.

Recall by (C.3) that for all $t \in [0,\infty)$, $\{Y'(t)=0\} \Rightarrow \{Y(t)=0\}$, thus,

$$\forall t \in [0,\infty) : \Pr(Y(t)=1, Y'(t)=0) = 0\,.$$

We divide our event space, therefore, into three complementary disjoint events: The first event is $Y(T_i)=Y'(T_i)=1$, the second is $0=Y(T_i)\neq Y'(T_i)=1$, and the third is $Y(T_i)=Y'(T_i)=0$. We shall show that all these three possibilities satisfy $D_i \leq D'_i$.

(a) If $Y(T_i)=Y'(T_i)=1$ then, with the fact that $c_s=0$, the social planer in Ω is indifferent between i sensing and i not sensing as both actions result in the same outcome, thus $D_i=0$. In a similar manner, since $Y'(T_i)=1$ in Ω', $D'_i=0$.

(b) If $Y'(T_i)=1$ and $Y(T_i)=0$, as explained above, $D'_i=0$. Regarding system Ω, when customer i senses there is the possibility that i will complete service in S_L and $X(T_i)+1$ customers (that is $X(T_i)$ customers in the queue plus customer i) would save the expense incurred by i joining the queue upon arrival. Thus the social cost when i senses is less than when he/she does not sense and $D_i \leq 0$.

(c) If $Y(T_i)=Y'(T_i)=0$) then we separate between two cases:

- When $\tau_i \leq T_{i+1}-T_i$, customer i completes service in S_L (and in S'_L) before the arrival of the next customer. Denote by B_i and B'_i the sojourn time of i in the queue of Ω and of Ω' respectively, measured in units of $\frac{1}{\mu}$. From property (e) of Lemma C.2 it is clear that $B'_i \leq B_i$. Since $\frac{c_w}{\mu}=1$, i sensing in Ω in this case saves $X(T_i)B_i+B_i-1$ relative to not sensing (that is $X(T_i)$ customers in the system saving B_i each, plus customer i saving B_i minus the cost of his/her own service). Likewise, i sensing in Ω' saves $X'(T_i)B'_i+B'_i-1$. Recall by (C.4) that $X'(T_i)\leq X(T_i)$, then we have:

$$D_i = 1-B_i-X(T_i)B_i \leq 1-B'_i-X'(T_i)B'_i = D'_i\,.$$

- When $\tau_i > T_{i+1}-T_i$, customer i, when sensing, begins service in S_L (S'_L) but with assumption (i) he/she is forced to join the queue upon the arrival of $i+1$. Denote by L and L' the number of customers that has left the system during the time period $[T_i, T_{i+1})$ in Ω and in Ω' respectively. Since $X'(T_i)\leq X(T_i)$, $L'\leq L$. Denote by B_i and B'_i the sojourn time of i in the queue of Ω and

of Ω' respectively, measured in units of $\frac{1}{\mu}$. As explained, $B'_i \leq B_i$. Then we have L customers in Ω saving the expense of waiting B_i each one and L' customers in Ω' saving the expense of waiting B'_i each one, therefore,

$$D_i = -LB_i \leq -L'B'_i = D'_i \,.$$

In summary, for all i such that $u_i > p'$, $D_i \leq D'_i$. Subsequently, if $J \subset \mathbb{N}$ is a subset of customers such that for all $j \in J$, $u_j > p'$, then

$$\sum_{j \in J} D_j \leq \sum_{j \in J} D'_j \,. \tag{C.17}$$

Assume now that each customer j in Ω senses if and only if $u_j \in [0, p) \cup [p', p' + \delta)$. This implies that the sensing probability in Ω is $p + \delta$. Similarly, if each customer in Ω' senses if and only if $u_j \in [0, p' + \delta)$, then the sensing probability in Ω' is $p' + \delta$. Define the set of customers $J := \{j \in \{1, \ldots, n\} \text{ s.t. } u_j \in [p', p' + \delta)\}$ then

$$F(p + \delta) - F(p) = \lim_{n \to \infty} \frac{1}{n} \sum_{j \in J} D_j \leq \lim_{n \to \infty} \frac{1}{n} \sum_{j \in J} D'_j = F(p' + \delta) - F(p') \,,$$

where the second inequality follows from (C.17). Therefore, by (C.16) we conclude that $D(p) \leq D(p')$ as desired. □

Since $D(p)$ is non-decreasing in p for all $\delta > 0$, we have that $\lim_{\delta \to 0} \frac{D(p)}{\delta}$ is also non-decreasing in p. This implies that $(\frac{d}{dp})F(p)$ is non-decreasing, and therefore $F(p)$ is convex in p. □

C.8 Proof of Proposition 4.8

Proof.

(a) Suppose $p = 0$, then the system is an M/M/1 queue with arrival rate Λ. This is socially optimal if and only if society prefers to bear all the social cost imposed by an individual that chooses not to sense rather than pay the cost of sensing. We tag the "transparent" customer as the one assigned the lowest priority, that is to say preempted by every other customer. Note that if this tagged customer senses, he/she will find S_L idle and no other customer will preempt him/her because $p = 0$. It was shown in Hassin and Haviv (2003) that since the arrival and action of the transparent customer imposes no externalities on other customers, the optimal

strategy for the transparent customer coincides with the socially optimal strategy. Thus, $p^* = 0$ if and only if the transparent customer prefers not to sense. As explained by Haviv and Ritov (1998), in an M/M/1 system, the queueing time of the transparent customer equals $\frac{1}{(\mu(1-\rho)^2)} - \frac{1}{\mu}$. This is the product of the expected queue length ($\frac{1}{(1-\rho)}$) and the expected time it takes to reduce the queue size by one ($\frac{1}{(\mu(1-\rho))}$) minus the expected service duration ($\frac{1}{\mu}$). We deduce that $p^* = 0$ if and only if:

$$c_w \left(\frac{1}{\mu(1-\rho)^2} - \frac{1}{\mu} \right) \leq c_s \quad \Leftrightarrow \quad \gamma \left(\frac{1}{(1-\rho)^2} - 1 \right) \leq 1 \quad \Leftrightarrow \quad \rho \leq 1 - \sqrt{\frac{\gamma}{1+\gamma}} \, .$$

(b) This result can be verified both algebraically and intuitively. Recall Proposition 4.4:

$$p_e = 0 \quad \Leftrightarrow \quad \rho \leq \frac{1}{1+\gamma} \, .$$

Note that:

$$0 \leq 1 - \sqrt{\frac{\gamma}{1+\gamma}} \leq \left(1 + \sqrt{\frac{\gamma}{1+\gamma}} \right) \left(1 - \sqrt{\frac{\gamma}{1+\gamma}} \right) = \frac{1}{1+\gamma} \, , \tag{C.18}$$

concluding that if $\rho \leq 1 - \sqrt{\frac{\gamma}{(1+\gamma)}}$ then $\rho \leq \frac{1}{(1+\gamma)}$.

We now give an intuitive explanation: Denote $W^0 := \mu^{-1} \cdot \frac{\rho}{(1-\rho)}$ the expected waiting time of a customer when $p = 0$ (as in the M/M/1 model). Suppose $p_e \neq 0$, then by Proposition 4.4, $\gamma \cdot \frac{\rho}{(1-\rho)} > 1$, which is equivalent to $c_w W^0 > c_s$. This means that an arbitrary customer can reduce both his/her own cost and the waiting cost of the others by sensing. Thus, $p^* \neq 0$, as claimed.

(c) By (C.18), if $\gamma > 0$ then

$$1 - \sqrt{\frac{\gamma}{1+\gamma}} < \frac{1}{1+\gamma} \, .$$

Taking $\rho \in [1 - \sqrt{\frac{\gamma}{(1+\gamma)}}, \frac{1}{(1+\gamma)}]$ implies by Proposition 4.4 and Proposition 4.8(a) that $p_e = 0 < p^*$.

□

C.9 Proof of Proposition 4.9

Proof. Denote $F'(1) := (\frac{d}{dp})F(p)|_{p=1}$ and $C'(1) := (\frac{d}{dp})C(p)|_{p=1}$. As discussed, $C(p)$ is convex, so,

$$p^* < 1 \quad \Leftrightarrow \quad 0 < C'(1) = \frac{d}{dp}(\frac{p}{\gamma} + F(p))\Big|_{p=1} \quad \Leftrightarrow \quad -\frac{1}{\gamma} < F'(1) \, .$$

Since $F(p)$ is convex we have that for all $h \in (0,1)$:

$$\frac{F(1) - F(1-h)}{h} < F'(1)\,.$$

Suppose that $h \in (0,1)$ satisfies (4.17), then

$$-\frac{1}{\gamma} \leq \frac{F(1) - F(1-h)}{h} < F'(1)\,,$$

and therefore $p^* < 1$. □

Appendix D

Proofs for Chapter 5

D.1 Proof of Proposition 5.1

Proof. Let T be an equilibrium cumulative distribution of tips, with first and second moments ET and ET^2 respectively[1]. Recall that the sum of the tip and the social cost is

$$t + \kappa \int (\tau - t)^2 dT(\tau) = \kappa t^2 - (2\kappa ET - 1)t + \kappa ET^2, \tag{D.1}$$

which is a convex quadratic function in t centered at $ET - 1/(2\kappa)$. Since onetime customers' expected waiting cost is constant and does not depend on their tip amount, restricted to non-negative tipping, they will all choose to tip $t_o = \{ET - 1/(2\kappa)\}^+$. For a repeat customer, tipping $t \in [0, t_o)$ is sub optimal because such a tip will assign him with priority level (and hence, expected waiting cost) strictly worse than any onetime customer, and, will yield a higher cost in the sum of tip and social cost than that of tipping t_o. Thus, $\underline{t}_r \geq t_o$. In addition, if $\underline{t}_r > t_o$, then a repeat customer tipping $\underline{t}_r$ can reduce his/her cost incurred by tipping and social cost by tipping any amount $t \in (t_o, \underline{t}_r)$, without changing her expected waiting cost, in contradiction to the equilibrium assumption. Therefore, $t_o = \underline{t}_r$, proving the second item.

The proof of the first item is similar to that of Glazer and Hassin (1986): Assume on the contrary that the tipping distribution for repeat customers, T_r, contains an atom at some $\tilde{t}_r$. Thus, the probability that at an arbitrary time point, a repeat customer who tipped t_r is waiting in the queue is strictly positive. Then, consider a repeat customer who arrives at that particular moment. By tipping $\tilde{t}_r + \varepsilon$ for some infinitesimal amount $\varepsilon > 0$, this arriving customer will cut in line all those customers who tipped $\tilde{t}_r$, receiving a strictly positive, ε-independent gain in waiting cost. Yet, her cost incurred from tipping

[1] Recall that the tips must be non-negative and that the tip amount is bounded from above by U, thus the first and second moments exist.

and social cost will differ to that of a customer who tipped $\tilde{t}_r$ only by an infinitesimal amount. Thus, she will receives a strictly higher expected utility than a customer tipping exactly $\tilde{t}_r$, in contradiction to the equilibrium assumption. Therefore in equilibrium T_r is continuous. Moreover, T_r is strictly increasing: Assume on the contrary that there exist a pair $t'_r < t''_r$ such that no repeat customer tips an amount $t \in (t'_r, t''_r)$. Recall that t_o minimizes the sum of the tip and social cost, and $t_o \leq t'_r < t''_r$. Thus, a customer who tips t''_r can reduce his/her cost incurred by tipping and social cost by tipping any amount $t \in (t'_r, t''_r)$, without changing her expected waiting cost, in contradiction to the equilibrium assumption. Therefore T_r is continuous and strictly increasing, proving the first item.

As explained, since onetime customers in equilibrium tip the least amount $\underline{t}_r$ (thus, receiving the lowest priority level), a repeat customer can always tip infinitesimally more than $\underline{t}_r$, receiving strictly higher utility in equilibrium. Therefore, we immediately obtain the third item. □

D.2 Proof of Proposition 5.2

Proof. Assume that the total arrival rate to the system in equilibrium is λ, and let $\lambda_r = \min\{\lambda, \Lambda_r\}$ be the rate of repeat customers. Consider the repeat customers tipping distribution function, $T_r(t;\lambda)$: From Equation (D.1), for $\kappa = 0$, the tip paid by onetime customers, $\underline{t}$, satisfies $\underline{t} = 0$. Therefore by Proposition 5.1, the domain of $T_r(t;\lambda)$ can be written as $(0, \underline{t}(\lambda)]$ for some $\underline{t}(\lambda) > 0$.

By Kleinrock (1976), the waiting time for a repeat customer tipping $t \in (0, \bar{t}(\lambda)]$ is given by

$$W_r(t;\lambda) = \frac{1}{\mu} \frac{1}{(1-(1-T_r(t;\lambda))\frac{\lambda_r}{\mu})^2}. \tag{D.2}$$

Specifically, for a repeat customer tipping 0_+, the expected waiting cost is equal to the total cost which is given by $\frac{c}{\mu}\frac{1}{(1-\frac{\lambda_r}{\mu})^2}$. Since all repeat customers receive the same expected utility in equilibrium we have, for all $t \in (0, \bar{t}(\lambda)]$:

$$\frac{c}{\mu}\frac{1}{(1-\frac{\lambda_r}{\mu})^2} = t + cW_r(t;\lambda) = t + \frac{c}{\mu}\frac{1}{(1-(1-T_r(t;\lambda))\frac{\lambda_r}{\mu})^3}. \tag{D.3}$$

Isolating $T_r(t;\lambda)$ we obtain:

$$T_r(t;\lambda) = \frac{\mu}{\lambda_r}\frac{1}{\sqrt{\frac{1}{(1-\frac{\lambda_r}{\mu})^2} - \frac{\mu}{c}t}} + 1 - \frac{\mu}{\lambda_r}.$$

Since the waiting time of a customer tipping $\bar{t}(\lambda)$ includes only his/her service time, then, by substituting $t = \bar{t}(\lambda)$ with $W_r(\bar{t}(\lambda); \lambda) = \frac{1}{\mu}$ in Equation (D.3) and isolating $\bar{t}(\lambda)$ we immediately get

$$\bar{t}(\lambda) = \frac{c}{\mu}\left(\frac{1}{(1-\frac{\lambda_r}{\mu})^2} - 1\right).$$

Moreover, from Equation D.3 it can be seen that

$$W_r(t;\lambda) = \frac{1}{\mu(1-\frac{\lambda_r}{\mu})^2} - \frac{t}{c}.$$

Finally, since onetime customers tip 0 and all receive the same priority level and commence service only when no repeat customers are in the system, by Kleinrock (1976), we have

$$W_o(\lambda) = \frac{1}{\mu(1-\frac{\lambda}{\mu})(1-\frac{\lambda_r}{\mu})}.$$

□

D.3 Proof of Proposition 5.4

Proof. Suppose that $\Lambda_r \in (0, \mu - \sqrt{\frac{c\mu}{U}}]$. By proposition 5.2 we have that $W_r(0;\lambda)$ is nondecreasing in λ. Using the definition of $C(\lambda)$, we have, for every $\lambda \in (0, \Lambda_r)$,

$$C(\lambda) = cW_r(0;\lambda) \le cW_r(0;\Lambda_r) \le cW_r(0;\mu - \sqrt{\frac{c\mu}{U}}) = \frac{c}{\mu(1-\frac{\mu-\sqrt{\frac{c\mu}{U}}}{\mu})^2} = U.$$

Therefore in equilibrium, all repeat customers join with rate Λ_r and by Proposition 5.2 tip according to $T(t;\Lambda_r)$ over $(0, \bar{t}(\Lambda_r)]$.

Similarly, if $\Lambda_r \in (\mu - \sqrt{\frac{c\mu}{U}}, \mu)$, then for every $\lambda \in (\Lambda_r, \mu)$

$$C(\lambda) = cW_o(\lambda) \ge cW_r(0;\lambda) \ge cW_r(0;\Lambda_r) > cW_r(0;\mu - \sqrt{\frac{c\mu}{U}}) = U.$$

Therefore in equilibrium, only repeat customers join with rate $\mu - \sqrt{c\mu/U} < \Lambda_r$ and tip according to $T_r(t;\mu - \sqrt{c\mu/U})$, and since $\bar{t}(\mu - \sqrt{c\mu/U}) = U - c/\mu$, the tipping domain is $(0, U - c/\mu]$. □

D.4 Proof of Proposition 5.7

Proof. Suppose the $T(t;m,\lambda)$ and $T_r(t;m,\lambda)$ are the equilibrium unconditional and repeat customers' tipping distributions, with a given (unconditional) mean tip m and

arrival rate λ, respectively. Let $\lambda_r = \min\{\lambda, \Lambda_r\}$ be the rate of repeat customers. As in the proof of Proposition 5.7, we have that

$$W_r(t; m, \lambda) = \frac{1}{\mu} \frac{1}{(1 - (1 - T_r(t; m, \lambda))\frac{\lambda_r}{\mu})^2}.$$

Suppose that the tipping domain is $[\underline{t}, \bar{t}]$. Since repeat customers are indifferent in equilibrium between any tip amount, we have that the total cost of a repeat customer tipping $\underline{t}_+$ is equal to that of a one tipping $t \in (\underline{t}, \bar{t}]$. Equivalently, for all $t \in (\underline{t}, \bar{t}]$,

$$\underline{t} + \frac{c}{\mu} \frac{1}{(1 - \frac{\lambda_r}{\mu})^2} + \kappa \int (\underline{t} - \tau)^2 dT(\tau; m, \lambda) = t + \frac{c}{\mu} \frac{1}{(1 - (1 - T_r(t; m, \lambda))\frac{\lambda_r}{\mu})^2} + \kappa \int (t - \tau)^2 dT(\tau; m, \lambda). \tag{D.4}$$

Denote by ET_2 the second moment of the distribution $T(t; m, \lambda)$. Assuming that the mean tip is m, we have that

$$\int (t - \tau)^2 dT(\tau; m, \lambda) = t^2 - 2mt + ET_2, \quad \forall t \in [\underline{t}, \bar{t}],$$

and therefore, by Equation (D.4), we have, for all $t \in (\underline{t}, \bar{t}]$,

$$\frac{c}{\mu} \frac{1}{(1 - \frac{\lambda_r}{\mu})^2} + \kappa \underline{t}^2 - (2\kappa m - 1)\underline{t} = \frac{c}{\mu} \frac{1}{(1 - (1 - T_r(t; m, \lambda))\frac{\lambda_r}{\mu})^2} + \kappa t^2 - (2\kappa m - 1)t. \tag{D.5}$$

Isolating $T_r(t; m, \lambda)$ in (D.5), we obtain

$$T_r(t; m, \lambda) = \frac{\mu}{\lambda_r} \frac{1}{\sqrt{\frac{1}{(1 - \frac{\lambda_r}{\mu})^2} - \frac{\mu\kappa}{c}(t + \underline{t} - 2(m - \frac{1}{2\kappa}))(t - \underline{t})}} + 1 - \frac{\mu}{\lambda_r}. \tag{D.6}$$

Note that $-\frac{\mu\kappa}{c}(t + \underline{t} - 2(m - \frac{1}{2\kappa}))(t - \underline{t})$ is concave decreasing for $t \geq \underline{t}$, and since $\frac{1}{\sqrt{x}}$ is convex and decreasing, we have that $T_r(t; m, \lambda)$ is convex increasing.

Similarly, isolating the expected waiting time in (D.5),

$$W_r(t; m, \lambda) = \frac{1}{\mu(1 - \frac{\lambda_r}{\mu})^2} - \frac{\kappa}{c}(t + \underline{t} - 2(m - \frac{1}{2\kappa}))(t - \underline{t}). \tag{D.7}$$

Since the waiting time of a customer tipping $\bar{t}$ includes only his/her service time, then, by substituting $t = \bar{t}$ in Equation (D.5) and rearranging we get a quadratic equation in $\bar{t}$

$$\bar{t}^2 - 2(m - \frac{1}{2\kappa})\bar{t} - \underline{t}(\underline{t} - 2(m - \frac{1}{2\kappa})) - \frac{c}{\kappa\mu}\left\{\frac{1}{(1 - \frac{\lambda_r}{\mu})^2} - 1\right\} = 0,$$

from which we solve for $\underline{t}$. Restricted to positive tipping we arrive at

$$\bar{t}(m,\lambda) = m - \frac{1}{2\kappa} + \sqrt{(m-\frac{1}{2\kappa})^2 + \frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\lambda_r}{\mu})^2} - 1\right\} + \underline{t}(\underline{t} - 2(m-\frac{1}{2\kappa}))}. \quad \text{(D.8)}$$

Finally, since the mean tip is m, we have from Equation D.1, that the tip payed by onetime customers, $\underline{t}$, as a function of m, satisfies $\underline{t}(m) = \{m - \frac{1}{2\kappa}\}^+$. Substituting $\underline{t} = \{m - \frac{1}{2\kappa}\}^+$ and $\lambda_r = \max\{\lambda, \Lambda_r\}$ in Equations (D.6), (D.7) and (D.8), we conclude that

$$T_r(t;m,\lambda) = \begin{cases} \frac{\mu}{\min\{\lambda,\Lambda_r\}} \frac{1}{\sqrt{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\mu\kappa}{c}(t^2 - 2(m-\frac{1}{2\kappa})t)}} + 1 - \frac{\mu}{\min\{\lambda,\Lambda_r\}}, & m < \frac{1}{2\kappa} \\ \frac{\mu}{\min\{\lambda,\Lambda_r\}} \frac{1}{\sqrt{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\mu\kappa}{c}(t-(m-\frac{1}{2\kappa}))^2}} + 1 - \frac{\mu}{\min\{\lambda,\Lambda_r\}}, & m > \frac{1}{2\kappa} \end{cases},$$

$$W_r(t;m,\lambda) = \frac{1}{\mu(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - \frac{\kappa}{c}\begin{cases} (t^2 - 2(m-\frac{1}{2\kappa})t), & m < \frac{1}{2\kappa} \\ (t-(m-\frac{1}{2\kappa}))^2, & m > \frac{1}{2\kappa} \end{cases}$$

and

$$\bar{t}(m,\lambda) = \begin{cases} m - \frac{1}{2\kappa} + \sqrt{(m-\frac{1}{2\kappa})^2 + \frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - 1\right\}}, & m < \frac{1}{2\kappa} \\ m - \frac{1}{2\kappa} + \sqrt{\frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\min\{\lambda,\Lambda_r\}}{\mu})^2} - 1\right\}}, & m > \frac{1}{2\kappa} \end{cases}$$

as desired. □

D.5 Proof of Proposition 5.8

Proof. Assume that the arrival rate to the system, λ is given, and let $\lambda_r = \min\{\lambda, \Lambda_r\}$ be the rate of repeat customers. Suppose that $m \le \frac{1}{2\kappa}$, so that $\underline{t}(m) = 0$. Thus, by definition, $ET(m,\lambda) = \frac{\lambda_r}{\lambda} ET_r(m,\lambda)$. To prove the first item, it suffices to show that $ET_r(m,\lambda)$ is convex increasing and positive in m. Let $t_p(m,\lambda)$ denote the the p-quantile tip amount of repeat customers' tipping distribution, i.e, t_p satisfies $T_r(t_p;m,\lambda) = p$. Substituting $\underline{t} = 0$ and $t = t_p$ in Equation (D.5) yields a quadratic equation in t_p whose positive solution is:

$$t_p(m,\lambda) = m - \frac{1}{2\kappa} + \sqrt{(m-\frac{1}{2\kappa})^2 + \frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\lambda_r}{\mu})^2} - \frac{1}{(1-(1-p)\frac{\lambda_r}{\mu})^2}\right\}}. \quad \text{(D.9)}$$

Since $x + \sqrt{x^2 + a}$ (with $a > 0$ a constant) is convex increasing and positive in $x > 0$, $t_p(m, \lambda)$ is convex increasing and positive as a function of $m \in [0, \frac{1}{2\kappa}]$, for every $p \in [0, 1]$. Note that

$$ET_r(m, \lambda) = \int_{p=0}^{1} t_p(m, \lambda) dp$$

and therefore $ET_r(m, \lambda)$ is also convex increasing and positive as a function of $m \in [0, \frac{1}{2\kappa}]$.

Suppose that $m > \frac{1}{2\kappa}$, such that $\underline{t}(m) = m - \frac{1}{2\kappa} > 0$. Substituting $\underline{t} = m - \frac{1}{2\kappa}$ and $t = t_p$ in Equation (D.5) yields a quadratic equation in t_p whose positive solution is:

$$t_p(m, \lambda) = m - \frac{1}{2\kappa} + \sqrt{\frac{c}{\kappa\mu}\left\{\frac{1}{(1 - \frac{\lambda_r}{\mu})^2} - \frac{1}{(1 - (1-p)\frac{\lambda_r}{\mu})^2}\right\}},$$

which is linear in m. Therefore,

$$\begin{aligned} ET_r(m, \lambda) &= \int_{p=0}^{1} t_p(m, \lambda) dp = \int_{p=0}^{1} m - \frac{1}{2\kappa} + \sqrt{\frac{c}{\kappa\mu}\left\{\frac{1}{(1 - \frac{\lambda_r}{\mu})^2} - \frac{1}{(1 - (1-p)\frac{\lambda_r}{\mu})^2}\right\}} dp \\ &= m - \frac{1}{2\kappa} + \int_{p=0}^{1} t_p(\frac{1}{2\kappa}, \lambda) dp = m - \frac{1}{2\kappa} + ET_r(\frac{1}{2\kappa}, \lambda). \end{aligned}$$

Thus, we have, for $m > \frac{1}{2\kappa}$

$$\begin{aligned} ET(m, \lambda) &= \frac{\lambda_r}{\lambda} ET_r(m, \lambda) + (1 - \frac{\lambda_r}{\lambda})\underline{t}(m) = \frac{\lambda_r}{\lambda}(m - \frac{1}{2\kappa} + ET_r(\frac{1}{2\kappa}, \lambda)) + (1 - \frac{\lambda_r}{\lambda})(m - \frac{1}{2\kappa}) \\ &= m - \frac{1}{2\kappa} + \frac{\lambda_r}{\lambda} ET_r(\frac{1}{2\kappa}, \lambda) = m - \frac{1}{2\kappa} + ET(\frac{1}{2\kappa}, \lambda). \end{aligned}$$

□

D.6 Proof of Proposition 5.9

Proof. First, we note that $\lim_{\rho \to 0} \Delta(\rho) = -\infty$, and $\lim_{\rho \to 1} \Delta(\rho) = \infty$. In addition, by basic algebraic manipulations,

$$\frac{d\Delta}{d\rho} = \frac{\frac{\rho\sqrt{\rho}}{\sqrt{2-\rho}(1-\rho)^2} + \frac{\pi}{2} - \arctan(\frac{1-\rho}{\sqrt{2\rho - \rho^2}})}{\rho^2}. \tag{D.10}$$

Since $\arctan(x) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ for all $x \in \mathcal{R}$, $\frac{d\Delta}{d\rho} > 0$ for all $\rho \in (0, 1)$, and therefore $\Delta(\rho)$ is strictly increasing in $\rho \in (0, 1)$. Thus, for every $\kappa > 0$ there exists a unique solution $\hat{\lambda}_r$ such that $\frac{1}{2\kappa} = \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu})$. Equivalently, $\hat{\lambda}_r = \mu\Delta^{-1}(\frac{1}{2}\sqrt{\frac{\mu}{c\kappa}})$, where Δ^{-1} is the inverse

of Δ, and since $\frac{1}{\sqrt{\kappa}}$ is decreasing in κ and Δ^{-1} is an increasing function, $\hat{\lambda}_r$ decreases in $\kappa > 0$.

Suppose that $\Lambda_r \geq \hat{\lambda}_r$. Since Δ is monotone increasing,

$$\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu}) \geq \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu}) = \frac{1}{2\kappa} > 0$$

and since $\frac{1}{\lambda}$ is decreasing and approaching 0 as $\lambda \to \infty$, there exists a unique solution $\hat{\lambda}_o \geq \Lambda_r$, such that $\frac{1}{2\kappa} = \frac{\Lambda_r}{\hat{\lambda}_o}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu})$.

By Proposition 5.8, $ET(m,\lambda)$ is convex increasing and positive in m, with slope equals 1 when $m \geq \frac{1}{2\kappa}$. Therefore, if λ is such that $\frac{1}{2\kappa} = ET(\frac{1}{2\kappa},\lambda)$, then, for every $m \geq \frac{1}{2\kappa}$, $m = ET(m,\lambda)$, and for every $m < \frac{1}{2\kappa}$, $m < ET(m,\lambda)$. To show that $\mathcal{M}(\lambda) = [\frac{1}{2\kappa}, +\infty)$, it suffices therefore to show that $\frac{1}{2\kappa} = ET(\frac{1}{2\kappa},\lambda)$.

Assume now that $\Lambda_r \geq \hat{\lambda}_r$, and recall the definition of $t_p(m,\lambda)$ in Equation (D.9). Then

$$\begin{aligned} ET(\frac{1}{2\kappa},\hat{\lambda}_r) &= \frac{\min\{\hat{\lambda}_r,\Lambda_r\}}{\hat{\lambda}_r}ET_r(\frac{1}{2\kappa},\hat{\lambda}_r) = \int_{p=0}^{1} t_p(\frac{1}{2\kappa},\hat{\lambda}_r)dp \\ &= \int_{p=0}^{1}\sqrt{\frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\hat{\lambda}_r}{\mu})^2} - \frac{1}{(1-(1-p)\frac{\hat{\lambda}_r}{\mu})^2}\right\}}dp = \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu}) = \frac{1}{2\kappa} \end{aligned} \tag{D.11}$$

where the second to last equality follows basic integration, and the last from the definition of $\hat{\lambda}_r$. Therefore, $\mathcal{M}(\hat{\lambda}_r) = [\frac{1}{2\kappa}, +\infty)$. Similarly, with $\hat{\lambda}_o \geq \Lambda_r$,

$$\begin{aligned} ET(\frac{1}{2\kappa},\hat{\lambda}_o) &= \frac{\min\{\hat{\lambda}_o,\Lambda_r\}}{\hat{\lambda}_o}ET(\frac{1}{2\kappa},\hat{\lambda}_o) = \frac{\Lambda_r}{\hat{\lambda}_o}\int_{p=0}^{1} t_p(\frac{1}{2\kappa},\hat{\lambda}_o)dp \\ &= \frac{\Lambda_r}{\hat{\lambda}_o}\int_{p=0}^{1}\sqrt{\frac{c}{\kappa\mu}\left\{\frac{1}{(1-\frac{\Lambda_r}{\mu})^2} - \frac{1}{(1-(1-p)\frac{\Lambda_r}{\mu})^2}\right\}}dp = \frac{\Lambda_r}{\hat{\lambda}_o}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu}) = \frac{1}{2\kappa}, \end{aligned} \tag{D.12}$$

and we conclude that $\mathcal{M}(\hat{\lambda}_r) = \mathcal{M}(\hat{\lambda}_o) = [\frac{1}{2\kappa}, +\infty)$. □

D.7 Proof of Lemma 5.10

Proof. Assume that the equilibrium arrival rate is $\lambda \in (0, \Lambda_r]$ such that only repeat customers join. Prior to proving the result of the lemma we show the following properties that we shall make use of in the proof:

1. $ET(m,\lambda)$ is increasing both in m and in λ and convex in m: Let $T_r(t;m,\lambda)$ be the tipping distribution of repeat customers and denote its k-th moment by $ET_r^k(m,\lambda)$. Note that since we focus here on equilibrium with only repeat customers joining (i.e., $\lambda \in (0,\Lambda_r]$), $T_r(t;m,\lambda)$ is also the unconditional tipping distribution.

   Suppose first that $m \leq \frac{1}{2\kappa}$, and recall the definition of $t_p(m,\lambda)$ from Equation (D.9). Since $T_r(t;m,\lambda)$ is defined over non-negatives and $\lambda_r = \lambda \in (0,\Lambda_r]$, we have

$$\begin{aligned} ET_r^k(m,\lambda) &= \int_{p=0}^{1} (t_p(m,\lambda))^k dp \\ &= \int_{p=0}^{1} \left(m - \frac{1}{2\kappa} + \sqrt{(m-\frac{1}{2\kappa})^2 + \frac{c}{\kappa\mu}\left\{ \frac{1}{(1-\frac{\lambda_r}{\mu})^2} - \frac{1}{(1-(1-p)\frac{\lambda_r}{\mu})^2} \right\}} \right)^k dp. \end{aligned}$$

   As explained, $t_p(m,\lambda)$ is positive increasing and convex in m for every p. In addition, by taking derivative with respect to λ we have that the term $\frac{1}{(1-\frac{\lambda}{\mu})^2} - \frac{1}{(1-(1-p)\frac{\lambda}{\mu})^2}$ is positive increasing in λ for every p, thus $(t_p(m,\lambda))^k$ is increasing in both coordinates for every p and $k > 0$. Therefore, $ET_r^k(m,\lambda)$ is increasing in both coordinates, and convex in m for every $k > 0$. Finally, since $\lambda \in (0,\Lambda_r]$, $ET^k(m,\lambda) = ET_r^k(m,\lambda)$, where $ET^k(m,\lambda)$ is the unconditional k-th moment, is increasing in both coordinates and convex in m, for every $k > 0$ (in particular, when $k = 1$).

   For $m \leq \frac{1}{2\kappa}$, we have that $T_r(t;m,\lambda) \sim m - \frac{1}{2\kappa} + T_r(t;\frac{1}{2\kappa},\lambda)$, and clearly $ET^k(m,\lambda)$ is increasing in m for every $k > 0$. In addition, since $ET^k(\frac{1}{2\kappa},\lambda)$ is increasing in λ for every $k > 0$, so is $ET^k(m,\lambda)$, in particular, when $k = 1$.

2. $C(m,\lambda)$ is increasing both in m and in λ: Suppose that $m \leq \frac{1}{2\kappa}$, so that $\underline{t}(m) = 0$. Assuming that $\lambda \in (0,\Lambda_r]$, we have,

$$C(m,\lambda) = \kappa \int \tau^2 dT_r(\tau;m,\lambda) + cW_r(0;m,\lambda) = \kappa ET_r^2(m,\lambda) + \frac{c}{\mu(1-\frac{\lambda}{\mu})^2}.$$

   From item 1 we know that $ET_r^2(m,\lambda)$ is increasing both in λ and m, and since $\frac{c}{\mu(1-\frac{\lambda}{\mu})^2}$ is in increasing in λ, we have that $C(m,\lambda)$ is increasing both in λ and m. If $m > \frac{1}{2\kappa}$, then $C(m,\lambda) = m - \frac{1}{2\kappa} + C(\frac{1}{2\kappa},\lambda)$, and since $C(\frac{1}{2\kappa},\lambda)$ is increasing in λ, $C(m,\lambda)$ is increasing both in λ and m.

We turn now to proving the result of the lemma: By Proposition 5.8, $ET(m,\lambda)$ is convex increasing and positive in m, with slope equals 1 for $m \geq \frac{1}{2\kappa}$. Therefore, for any $\lambda \in (0,\Lambda_r]$, exactly one of the following options holds:

- If $ET(\frac{1}{2\kappa}, \lambda) < \frac{1}{2\kappa}$, then $\mathcal{M}(\lambda) = \{m_\lambda\}$ is a singleton and satisfies $m_\lambda < \frac{1}{2\kappa}$.
- If $ET(\frac{1}{2\kappa}, \lambda) = \frac{1}{2\kappa}$, then $\mathcal{M}(\lambda) = [\frac{1}{2\kappa}, \infty)$.
- If $ET(\frac{1}{2\kappa}, \lambda) > \frac{1}{2\kappa}$, then $\mathcal{M}(\lambda) = \emptyset$.

Recall from Equation (D.11) that for $\lambda < \Lambda_r$, $ET(\frac{1}{2\kappa}, \lambda) = \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\lambda}{\mu})$, which, as explained in the proof of Proposition 5.9, is an increasing function in λ. Since $\hat{\lambda}_r$ by definition satisfies that $\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu}) = \frac{1}{2\kappa}$, then $ET(\frac{1}{2\kappa}, \lambda) \leq \frac{1}{2\kappa}$ if and only if $\lambda \leq \hat{\lambda}_r$ (with equality if and only if $\lambda = \hat{\lambda}_r$). Thus, we obtain

$$\mathcal{M}(\lambda) = \begin{cases} \{m_\lambda\} \text{ for some } m_\lambda < \frac{1}{2\kappa} & \text{if } \lambda < \hat{\lambda}_r, \\ [\frac{1}{2\kappa}, \infty) & \text{if } \lambda = \hat{\lambda}_r, \\ \emptyset & \text{if } \lambda > \hat{\lambda}_r. \end{cases} \tag{D.13}$$

For $\lambda = \hat{\lambda}_r$, since $C(m, \lambda) = m - \frac{1}{2\kappa} + \hat{C}_r$ is linear increasing in $m \in [\frac{1}{2\kappa}, \infty)$, we have that $\mathcal{C}(\hat{\lambda}_r) = [\hat{C}_r, \infty)$.

Consider $\lambda \in (0, \hat{\lambda}_r)$, and denote by m_λ the unique solution $m_\lambda = ET(m_\lambda, \lambda)$. Let $g(m)$ denote $ET(m, \lambda)$ as a univariate function of m, and $g^{(n)}(m)$ its n-th function power. Thus, $g(m)$ is an increasing function, possessing a unique fixed point $m_\lambda = g(m_\lambda)$, and for every $m > 0$, $\lim_{n\to\infty} g^{(n)}(m) = m_\lambda$. We next show that $C(\lambda)$ is increasing in λ: Consider $\tilde{\lambda} < \lambda$, with $m_{\tilde{\lambda}} = ET(m_{\tilde{\lambda}}, \tilde{\lambda})$. Then, since ET is increasing in each coordinate,

$$m_{\tilde{\lambda}} = ET(m_{\tilde{\lambda}}, \tilde{\lambda}) < ET(m_{\tilde{\lambda}}, \lambda) = g(m_{\tilde{\lambda}})$$

and by induction, $m_{\tilde{\lambda}} < g^{(n)}(m_{\tilde{\lambda}})$ for every $n \in \mathbb{N}$. Therefore, $m_{\tilde{\lambda}} < \lim_{n\to\infty} g^{(n)}(m_{\tilde{\lambda}}) = m_\lambda$. Thus, with item 2, we get

$$C(\tilde{\lambda}) = C(m_{\tilde{\lambda}}, \tilde{\lambda}) < C(m_\lambda, \lambda) = C(\lambda).$$

Combining all together we obtain

$$\mathcal{C}(\lambda) = \begin{cases} \{C(\lambda)\} & \text{if } \lambda < \hat{\lambda}_r, \\ [\hat{C}_r, \infty) & \text{if } \lambda = \hat{\lambda}_r, \\ \emptyset & \text{if } \lambda > \hat{\lambda}_r, \end{cases}$$

and $C(\lambda)$ is increasing in $\lambda \in (0, \hat{\lambda}_r]$. □

D.8 Proof of Proposition 5.11

Proof. As explained, every arrival rate $\lambda \in (0, \Lambda_r]$ in equilibrium must satisfy $U \in \mathcal{C}(\lambda)$.

Suppose that $\Lambda_r < \hat{\lambda}_r$ and recall from Lemma 5.10 that for every $\lambda < \hat{\lambda}_r$, $\mathcal{C}(\lambda) = \{C(\lambda)\}$ is a singleton with $C(\lambda)$ increasing in λ. If $U > C(\Lambda_r)$, then, $U > C(\lambda)$ for every $\lambda \in (0, \Lambda_r]$, therefore $U \notin \mathcal{C}(\lambda)$ for every $\lambda \in (0, \Lambda_r]$ and an equilibrium with $\lambda \in (0, \Lambda_r]$ does not exist. If, however, $U \leq C(\Lambda_r)$ (note that $C(0) = 0$) then there exists a unique $\lambda \in (0, \Lambda_r]$ such that $U = C(\lambda)$ (equivalently, $U \in \mathcal{C}(\lambda)$). This value λ satisfies that $\mathcal{M} = \{m\}$ is a singleton with $m < \frac{1}{2\kappa}$.

Suppose on the contrary that $\Lambda_r \geq \hat{\lambda}_r$. If $U < \hat{C}_r = C(\hat{\lambda}_r)$ then again since $C(\lambda)$ is increasing over $(0, \hat{\lambda}_r]$, there exists a unique $\lambda \in (0, \hat{\lambda}_r)$ such that $U = C(\lambda)$ (equivalently, $U \in \mathcal{C}(\lambda)$) and that $\mathcal{M} = \{m\}$ is a singleton with $m < \frac{1}{2\kappa}$. If $U \geq \hat{C}_r$, then $U \in [\hat{C}_r, \infty) = \mathcal{C}(\hat{\lambda}_r)$ and for every $\lambda \neq \hat{\lambda}_r$, $U \notin \mathcal{C}(\lambda)$. Thus, there exists a unique equilibrium with rate $\hat{\lambda}_r$ and mean tip m such that

$$U = C(m, \hat{\lambda}_r) = m - \frac{1}{2\kappa} + C(\frac{1}{2\kappa}, \hat{\lambda}_r) = m - \frac{1}{2\kappa} + \hat{C}_r,$$

or, equivalently, $m = \frac{1}{2\kappa} + U - \hat{C}_r$. □

D.9 Proof of Lemma 5.12

Proof. Recall the definition of $\hat{\lambda}_r$ as the unique solution satisfying $\frac{1}{2\kappa} = \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu})$. When ρ approaches 1, $\Delta(\rho)$ grows unboundedly, therefore $\hat{\lambda}_r < \mu$. Thus, $\frac{1}{2\kappa} > \frac{\hat{\lambda}_r}{\mu} \cdot \frac{1}{2\kappa} = \frac{\hat{\lambda}_r}{\mu}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu})$, and since both $\Delta(\frac{\lambda}{\mu})$ and $\frac{\lambda}{\mu}$ are increasing and positive in $\lambda \in (0, \mu)$, there exists a unique solution $\hat{\Lambda}_r > \hat{\lambda}_r$ such that $\frac{1}{2\kappa} > \frac{\hat{\Lambda}_r}{\mu}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\Lambda}_r}{\mu})$.

Suppose that $\Lambda_r \in [\hat{\lambda}_r, \hat{\lambda}_r)$. Similarly to the proof of Lemma 5.10, we have that

- If $ET(\frac{1}{2\kappa}, \lambda) < \frac{1}{2\kappa}$, then $\mathcal{M}(\lambda) = \{m_\lambda\}$ is a singleton and satisfies $m_\lambda < \frac{1}{2\kappa}$.
- If $ET(\frac{1}{2\kappa}, \lambda) = \frac{1}{2\kappa}$, then $\mathcal{M}(\lambda) = [\frac{1}{2\kappa}, \infty)$.
- If $ET(\frac{1}{2\kappa}, \lambda) > \frac{1}{2\kappa}$, then $\mathcal{M}(\lambda) = \emptyset$.

Recall from Equation (D.12) that for $\lambda \in (\Lambda_r, \mu)$, $ET(\frac{1}{2\kappa}, \lambda) = \frac{\Lambda_r}{\lambda}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu})$, which, as explained in the proof of Proposition 5.9, is a decreasing function in λ. Since $\hat{\lambda}_o$ by

definition satisfies that $\frac{\Lambda_r}{\hat{\lambda}_o}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu}) = \frac{1}{2\kappa}$, then $ET(\frac{1}{2\kappa},\lambda) \le \frac{1}{2\kappa}$ if and only if $\lambda \ge \hat{\lambda}_r$ (with equality if and only if $\lambda = \hat{\lambda}_o$). Thus, we obtain

$$\mathcal{M}(\lambda) = \begin{cases} \{m_\lambda\} \text{ for some } m_\lambda < \frac{1}{2\kappa} & \text{if } \lambda > \hat{\lambda}_o, \\ [\frac{1}{2\kappa},\infty) & \text{if } \lambda = \hat{\lambda}_o, \\ \emptyset & \text{if } \lambda < \hat{\lambda}_o. \end{cases} \tag{D.14}$$

For $\lambda = \hat{\lambda}_r$, since $C(m,\lambda) = m - \frac{1}{2\kappa} + \hat{C}_r$ is linear increasing in $m \in [\frac{1}{2\kappa},\infty)$, we have that $\mathcal{C}(\hat{\lambda}_o) = [\hat{C}_o,\infty)$.

Consider $\lambda > \hat{\lambda}_o$, with $m_\lambda \in \mathcal{M}(\lambda)$ and $m_\lambda < \frac{1}{2\kappa}$. Note that in this case $\underline{t}(m_\lambda) = 0$ and

$$C(\lambda) = C(m_\lambda,\lambda) = \kappa\int \tau^2 dT(\tau;m_\lambda,\lambda) + cW_o(\lambda) = \kappa\frac{\Lambda_r}{\lambda}ET_r^2(m_\lambda,\Lambda_r) + \frac{c}{\mu(1-\frac{\lambda}{\mu})(1-\frac{\Lambda_r}{\mu})}. \tag{D.15}$$

We next show that $C(\lambda)$ is convex in $\lambda \in (\Lambda_r,\infty)$. To this purpose we first note that m_λ is convex decreasing with respect to λ: Since $m_\lambda \in \lambda$ and $\underline{t}(m_\lambda) = 0$, $m_\lambda = \frac{\Lambda_r}{\lambda}T_r(m_\lambda,\Lambda_r)$. Taking derivative with respect to λ we get

$$\frac{dm_\lambda}{d\lambda} = \frac{-\Lambda_r}{\lambda^2}T_r(m_\lambda,\Lambda_r) + \frac{\Lambda_r}{\lambda}\frac{dT_r(m_\lambda,\Lambda_r)}{dm_\lambda}\frac{dm_\lambda}{d\lambda} \quad \Rightarrow \quad \frac{dm_\lambda}{d\lambda} = -\frac{T_r(m_\lambda,\Lambda_r)}{\frac{\lambda^2}{\Lambda_r} - \lambda\frac{dT_r(m_\lambda,\Lambda_r)}{dm_\lambda}}. \tag{D.16}$$

Recall by Proposition 5.8 that $\frac{dT_r(m,\Lambda_r)}{m} \le 1$, thus, $\frac{dm_\lambda}{d\lambda}$ is negative and m_λ is decreasing in λ. Moreover, since $T_r(m,\Lambda_r)$ is convex in m, $\frac{dT_r(m_\lambda,\Lambda_r)}{m_\lambda}$ is increasing in m_λ and decreasing in λ, implying that $\frac{\lambda^2}{\Lambda_r} - \lambda\frac{dT_r(m_\lambda,\Lambda_r)}{dm_\lambda}$ is increasing with λ. $T_r(m_\lambda,\Lambda_r)$ is increasing with m_λ, thus, decreasing with λ. Overall, we have that $\frac{dm_\lambda}{d\lambda}$ is negative and monotone increasing concluding that m_λ is decreasing and convex in λ.

Observing the expression for $C(\lambda)$ in Equation (D.15), we note that from item 1 in the proof of Lemma 5.10, $ET_r^2(m_\lambda,\lambda)$ is positive convex increasing in m_λ, and since m_λ is convex decreasing, $ET_r^2(m_\lambda,\lambda)$ is positive convex decreasing in λ. Clearly, $\frac{1}{\lambda}$ is convex decreasing and positive therefore $\frac{1}{\lambda}ET_r^2(m_\lambda,\lambda)$ is positive convex decreasing in λ as the product of two such functions. We further note that $\frac{1}{1-\frac{\lambda}{\mu}}$ is convex increasing and positive, approaching ∞ as $\lambda \to \mu$. We conclude that $C(\lambda)$ is convex and positive in $\lambda \in (\hat{\lambda}_o,\mu)$, with $\lim_{\lambda\to\mu} C(\lambda) = \infty$.

Suppose that $\Lambda_r < \hat{\lambda}_r$, meaning that for every $\lambda \in (\Lambda_r,\mu)$,

$$ET(\frac{1}{2\kappa},\lambda) = \frac{\Lambda_r}{\lambda}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu}) < \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu}) < \sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu}) = \frac{1}{2\kappa}.$$

Then $\mathcal{M}(\lambda) = \{m_\lambda\}$ is a singleton and satisfies $m_\lambda < \frac{1}{2\kappa}$ for all $\lambda \in (\Lambda_r, \mu)$, and, as explained, $C(\lambda)$ is convex and positive with $\lim_{\lambda\to\mu} C(\lambda) = \infty$.

Finally, if $\Lambda_r > \hat{\Lambda}_r$, then for every $\lambda \in (\Lambda_r, \mu)$

$$ET(\frac{1}{2\kappa}, \lambda) = \frac{\Lambda_r}{\lambda}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\Lambda_r}{\mu}) > \frac{\hat{\Lambda}_r}{\lambda}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\Lambda}_r}{\mu}) > \frac{\hat{\Lambda}_r}{\mu}\sqrt{\frac{c}{\kappa\mu}}\Delta(\frac{\hat{\Lambda}_r}{\mu}) = \frac{1}{2\kappa},$$

which implies that $\mathcal{C}(\lambda) = \mathcal{M}(\lambda) = \emptyset$. □

D.10 Proof of Lemma 5.13

Proof. First, we derive a simplified expression for $C(\lambda)$ (Equation (D.18) below) on which we base our analysis in the proof: For a given equilibrium arrival rate $\lambda > \max\{\Lambda_r, \hat{\lambda}_o\}$, suppose that m_λ is the unique value satisfying $m_\lambda \in \mathcal{M}$ and $m_\lambda \leq \frac{1}{2\kappa}$ (by Proposition 5.8, such value must exist uniquely). Thus, $\underline{t}(m_\lambda) = 0$. In equilibrium, repeat customers are indifferent between tipping any amount in $(0, \bar{t}(m_\lambda, \lambda)]$. In particular, a repeat customer can choose an independent random tip $X \sim T_r(t; m_\lambda, \lambda)$. Note that the expected waiting time of that randomizing customer is, in fact, the average waiting time of repeat customers in the population, which is equal to $\frac{1}{\mu-\Lambda_r}$, and her expected tip is the average repeat customers' tip, which is equal to $ET_r(m_\lambda, \lambda) = \frac{\lambda}{\Lambda_r}m_\lambda$. Also note that $ET_r^2(m_\lambda, \lambda) = \frac{\lambda}{\Lambda_r}ET^2(m_\lambda, \lambda)$. The expected total cost for such customer is therefore given by

$$\begin{aligned}
&E(X) + \kappa E\left(\int (X-\tau)^2 dT_r(\tau; m_\lambda, \lambda)\right) + cE(W_r(t; m_\lambda, \lambda)) = \\
&E(X) + \kappa(E(X^2) - 2E(X)\cdot ET(m_\lambda, \lambda) + ET^2(m_\lambda, \lambda)) + \frac{c}{\mu - \Lambda_r} = \\
&\frac{\lambda}{\Lambda_r}m_\lambda + \kappa(1 + \frac{\lambda}{\Lambda_r})ET^2(m_\lambda, \lambda) - 2\kappa\frac{\lambda}{\Lambda_r}m_\lambda^2 + \frac{c}{\mu - \Lambda_r}.
\end{aligned}$$

This cost is equal to that of a repeat customer who tips 0^+, thus,

$$\frac{\lambda}{\Lambda_r}m_\lambda(1 - 2\kappa m_\lambda) + \kappa(1 + \frac{\lambda}{\Lambda_r})ET^2(m_\lambda, \lambda) + \frac{c}{\mu - \Lambda_r} = \kappa ET^2(m_\lambda, \lambda) + \frac{c}{\mu(1 - \frac{\Lambda_r}{\mu})^2}.$$

Upon rearranging we get an expression for $ET^2(m_\lambda, \lambda)$ (equivalently, $\frac{\Lambda_r}{\lambda}ET_r^2(m_\lambda, \lambda)$):

$$ET^2(m_\lambda, \lambda) = \frac{c\Lambda_r^2}{\kappa\lambda(\mu - \Lambda_r)^2} + 2m_\lambda(m_\lambda - \frac{1}{2\kappa}). \tag{D.17}$$

Substituting the latter expression in Equation (D.15), we have

$$C(\lambda) = \frac{c\Lambda_r^2}{(\mu-\Lambda_r)^2}\frac{1}{\lambda} + 2\kappa m_\lambda(m_\lambda - \frac{1}{2\kappa}) + \frac{c}{\mu(1-\frac{\lambda}{\mu})(1-\frac{\Lambda_r}{\mu})}. \tag{D.18}$$

Suppose now that $\Lambda_r > \hat{\lambda}_r$, so that $\hat{\lambda}_o > \Lambda_r$. From Lemma 5.12, $C(\lambda)$ is convex over $(\hat{\lambda}_o, \mu)$ and approaches ∞ when $\lambda \to \mu$. Therefore, $C(\lambda)$ is also monotone increasing over $(\hat{\lambda}_o, \mu)$ if and only if its right derivative evaluated at $\hat{\lambda}_o$ is nonnegative, otherwise $C(\lambda)$ is non-monotone. In other words, $\hat{C}_o = \underline{C}$ if and only if $\frac{dC}{d\lambda}|_{\lambda=\hat{\lambda}_o} \geq 0$. Thus, concerning $\Lambda_r > \hat{\lambda}_r$, it suffices to show that there exists $\overline{\Lambda}_r$ satisfying $\hat{\lambda}_r < \overline{\Lambda}_r < \hat{\Lambda}_r$, such that $\frac{dC}{d\lambda}|_{\lambda=\hat{\lambda}_o} < 0$ when $\Lambda_r \in (\hat{\lambda}_r, \overline{\Lambda}_r)$ and $\frac{dC}{d\lambda}|_{\lambda=\hat{\lambda}_o} \geq 0$ when $\Lambda_r \in (\overline{\Lambda}_r, \hat{\Lambda}_r)$.

Taking derivative of (D.18) with respect to λ and substituting $\lambda = \hat{\lambda}_o$ (note that $\lim_{\lambda\to\hat{\lambda}_o^+} m_\lambda = \frac{1}{2\kappa}$) we arrive at

$$\begin{aligned}
\left.\frac{dC}{d\lambda}\right|_{\lambda=\hat{\lambda}_o} &= \frac{c\mu}{(\mu-\Lambda_r)\left(\mu-\hat{\lambda}_o\right)^2} - c\left(\frac{\Lambda_r}{\hat{\lambda}_o(\mu-\Lambda_r)}\right)^2 + \left.\frac{dm_\lambda}{d\lambda}\right|_{\lambda=\hat{\lambda}_o} \\
&= \frac{c\mu}{(\mu-\Lambda_r)\left(\mu-\hat{\lambda}_o\right)^2} - c\left(\frac{\Lambda_r}{\hat{\lambda}_o(\mu-\Lambda_r)}\right)^2 - \frac{T_r(\frac{1}{2\kappa},\Lambda_r)}{\frac{\hat{\lambda}_o^2}{\Lambda_r} - \hat{\lambda}_o\left.\frac{dT_r(m,\Lambda_r)}{dm}\right|_{m=\frac{1}{2\kappa}}} \\
&= \frac{c\mu}{(\mu-\Lambda_r)\left(\mu-\hat{\lambda}_o\right)^2} - c\left(\frac{\Lambda_r}{\hat{\lambda}_o(\mu-\Lambda_r)}\right)^2 - \frac{1}{2\kappa(\hat{\lambda}_o-\Lambda_r)}.
\end{aligned} \tag{D.19}$$

where the second equality follows from Equation (D.16) and the third from the fact that $ET_r(\frac{1}{2\kappa},\Lambda_r) = \frac{\hat{\lambda}_o}{\Lambda_r}ET(\frac{1}{2\kappa},\hat{\lambda}_o) = \frac{\hat{\lambda}_o}{\Lambda_r}\cdot\frac{1}{2\kappa}$ and the fact that $\left.\frac{dT_r(m,\Lambda_r)}{dm}\right|_{m=\frac{1}{2\kappa}} = 1$ (see Proposition 5.8).

Consider the right-hand side of Equation (D.19) as a function of Λ_r. From the definition of $\hat{\lambda}_o$ we have

$$\hat{\lambda}_o - \Lambda_r = (2\sqrt{\frac{c\kappa}{\mu}}\Delta(\frac{\Lambda_r}{\mu}) - 1)\Lambda_r.$$

Note that $\Delta(\rho)$ is increasing. By assumption, together with the definition of $\hat{\lambda}_r$, we have $\Lambda_r > \hat{\lambda}_r = \mu\Delta^{-1}(\frac{1}{2}\sqrt{\frac{\mu}{c\kappa}})$, and therefore, $2\sqrt{\frac{c\kappa}{\mu}}\Delta(\frac{\Lambda_r}{\mu}) > 2\sqrt{\frac{c\kappa}{\mu}}\Delta(\frac{\hat{\lambda}_r}{\mu}) = 1$. Thus, $\hat{\lambda}_o - \Lambda_r$ is positive increasing in Λ_r, and so, $\hat{\lambda}_o$ is also positive increasing as a function of Λ_r, for $\Lambda_r \in (\hat{\lambda}_r, \mu)$. Hence, the term $\frac{-1}{\hat{\lambda}_o-\Lambda_r}$ is increasing in Λ_r, and so is the term $\frac{1}{\mu-\Lambda_r}\cdot\frac{1}{(\mu-\hat{\lambda}_o)^2}$ as it is a product of positive increasing functions of Λ_r. In addition, we have that

$$\frac{\Lambda_r}{\hat{\lambda}_o(\mu-\Lambda_r)} = \frac{1}{2}\sqrt{\frac{\mu}{c\kappa}}\cdot\frac{1}{(\mu-\Lambda_r)\Delta(\frac{\Lambda_r}{\mu})}$$

Simple algebraic manipulations, together with the definition of $\Delta(\rho)$ yield:

$$\frac{d}{d\Lambda_r}\left((\mu-\Lambda_r)\Delta(\frac{\Lambda_r}{\mu})\right)=\frac{\mu^2}{\Lambda_r^2}\left(\frac{\pi}{2}-\arctan\left(\frac{1-\frac{\Lambda_r}{\mu}}{\sqrt{\frac{\Lambda_r}{\mu}\left(2-\frac{\Lambda_r}{\mu}\right)}}\right)-\sqrt{\frac{\Lambda_r}{\mu}\left(2-\frac{\Lambda_r}{\mu}\right)}\right)>0$$

where the inequality is because the function

$$f(\rho)=\arctan\left(\frac{1-\rho}{\sqrt{\rho\,(2-\rho)}}\right)+\sqrt{\rho\,(2-\rho)}$$

is decreasing over $(0,1)$ with $\lim_{\rho\to 0^+}=\frac{\pi}{2}$. Therefore, $\frac{\Lambda_r}{\hat{\lambda}_o(\mu-\Lambda_r)}$ is decreasing and positive, and $-\left(\frac{\Lambda_r}{\hat{\lambda}_o(\mu-\Lambda_r)}\right)^2$ is increasing. Overall, we have that the term in (D.19) is a sum of increasing functions, therefore increasing in Λ_r for $\Lambda_r>\hat{\lambda}_r$. Note that when $\Lambda\to\hat{\lambda}_r$, then $\hat{\lambda}_o\to\Lambda_r$, and the term in (D.19) approaches $-\infty$, while when $\Lambda\to\hat{\Lambda}_r$, then $\hat{\lambda}_o\to\mu$, and the term in (D.19) approaches ∞. Thus, there exists some $\overline{\Lambda}_r$, such that $\frac{dC}{d\lambda}\big|_{\lambda=\hat{\lambda}_o}<0$ if $\Lambda_r\in(\hat{\lambda}_r,\overline{\Lambda}_r)$, and $\frac{dC}{d\lambda}\big|_{\lambda=\hat{\lambda}_o}\geq 0$ if $\Lambda_r\in(\overline{\Lambda}_r,\hat{\Lambda}_r)$.

Suppose now that $\Lambda_r<\hat{\lambda}_r$. Similarly to the previous case with $\Lambda_r>\hat{\lambda}_r$, here it suffices to show that there exist $\underline{\Lambda}'_r$ and $\underline{\Lambda}''_r$ satisfying $0<\underline{\Lambda}'_r<\underline{\Lambda}''_r<\hat{\lambda}_r$, such that $\frac{dC}{d\lambda}|_{\lambda=\Lambda_r}\geq 0$ when $\Lambda_r\in(0,\underline{\Lambda}'_r]$ and $\frac{dC}{d\lambda}|_{\lambda=\Lambda_r}<0$ when $\Lambda_r\in(\underline{\Lambda}''_r,\hat{\lambda}_r)$.

Taking derivative of (D.18) with respect to λ and substituting $\lambda=\Lambda_r$ we arrive at

$$\begin{aligned}\left.\frac{dC}{d\lambda}\right|_{\lambda=\Lambda_r}&=\frac{c\mu}{(\mu-\Lambda_r)^3}+\frac{c}{(\mu-\Lambda_r)^2}+(4\kappa m_{\Lambda_r}-1)\left.\frac{dm_\lambda}{d\lambda}\right|_{\lambda=\Lambda_r}\\&=\frac{c\mu}{(\mu-\Lambda_r)^3}+\frac{c}{(\mu-\Lambda_r)^2}+(4\kappa m_{\Lambda_r}-1)\frac{-T_r(m_{\Lambda_r},\Lambda_r)}{\Lambda_r-\Lambda_r\left.\frac{dT_r(m,\Lambda_r)}{dm}\right|_{m=m_{\Lambda_r}}}\\&=\frac{c\mu}{(\mu-\Lambda_r)^3}+\frac{c}{(\mu-\Lambda_r)^2}-\frac{m_{\Lambda_r}(4\kappa m_{\Lambda_r}-1)}{\Lambda_r(1-\left.\frac{dT_r(m,\Lambda_r)}{dm}\right|_{m=m_{\Lambda_r}})}\end{aligned}\tag{D.20}$$

where the second equality follows from Equation (D.16) and the third from the fact that $ET_r(m_{\Lambda_r},\Lambda_r)=ET(m_{\Lambda_r},\hat{\lambda}_o)=m_{\Lambda_r}$.

Consider Equation D.20 as a function of Λ_r. By Proposition 5.8, when $\Lambda_r<\hat{\lambda}_r$ we have $1>\left.\frac{dT_r(m,\Lambda_r)}{dm}\right|_{m=m_{\Lambda_r}}$. Note that when $\Lambda\to\hat{\lambda}_o^+$, $m_{\Lambda_r}\to\frac{1}{2\kappa}$ and $\left.\frac{dT_r(m,\Lambda_r)}{m}\right|_{m=m_{\Lambda_r}}$ approaches 1, therefore the term in (D.20) approaches $-\infty$. Thus there exists some $\underline{\Lambda}''_r<\hat{\lambda}_r$ such that $\frac{dC}{d\lambda}\big|_{\lambda=\Lambda_r}<0$ for all $\Lambda_r\in(\underline{\Lambda}''_r,\hat{\lambda}_r)$. When $\Lambda_r\to 0$, $m_{\Lambda_r}\to 0$, and since m_λ is decreasing for every $\lambda>\Lambda_r$, the term $(4\kappa m_{\Lambda_r}-1)\frac{dm}{d\lambda}$ approaches a nonnegative value. Thus, $\lim_{\Lambda_r\to 0}\left(\frac{dC}{d\lambda}\big|_{\lambda=\Lambda_r}\right)>0$, so there exists some $\underline{\Lambda}'_r\in(0,\underline{\Lambda}''_r)$ such that $\frac{dC}{d\lambda}\big|_{\lambda=\Lambda_r}<0$ for all $\Lambda_r\in(0,\underline{\Lambda}'_r]$. □

D.11 Proof of Proposition 5.14

Proof. As explained, every arrival rate $\lambda \in (0, \Lambda_r]$ in equilibrium must satisfy $U \in \mathcal{C}(\lambda)$. Recall from Lemma 5.12 that for every $\lambda \in (\Lambda_r, \mu)$, $\mathcal{C}(\lambda) = \{C(\lambda)\}$ is a singleton with $C(\lambda)$ convex in λ and approaches ∞ as λ goes to μ, and $\underline{\lambda} = \arg\min_{\lambda \in (\Lambda_r, \mu)} C(\lambda)$.

Suppose that $\Lambda_r < \hat{\lambda}_r$:

- If $U > C(\Lambda_r)$, then, $U > C(\lambda)$ (i.e., $U \notin \mathcal{C}(\lambda)$) for every $\lambda \in (0, \underline{\lambda}]$, and, there exists one $\lambda \in (\underline{\lambda}, \mu)$ such that $C(\lambda) = U$ (i.e., $U \in \mathcal{C}(\lambda)$). Therefore there exists an equilibrium with $\lambda \in (\underline{\lambda}, \mu)$ and mean tip $m < \frac{1}{2\kappa}$.
- If $U \in [\underline{C}, C(\Lambda_r)]$, then, since $C(\lambda)$ is convex, there exists one $\lambda_1 \in [\Lambda_r, \underline{\lambda}]$ such that $C(\lambda_1) = U$, and (in case $\underline{C} < C(\Lambda_r)$ another $\lambda_2 \in (\underline{\lambda}, \mu)$ such that $C(\lambda_2) = U$. Therefore there exist two equilibria, one with $\lambda_1 \in [\Lambda_r, \underline{\lambda}]$ and mean tip m_1, and another one with $\lambda_2 \in (\underline{\lambda}, \mu)$ and mean tip m_2, such that $m_1, m_2 < \frac{1}{2\kappa}$.
- If $U < \underline{C}$ then for every $\lambda \in (\Lambda_r, \mu)$, $U < C(\lambda)$ implying that an equilibrium with $\lambda \in (\Lambda_r, \mu)$ does not exist.

Suppose that $\Lambda_r \in [\hat{\lambda}_r, \hat{\Lambda}_r)$:

- If $U > \hat{C}_o$, then by Lemma 5.12, $U \in \mathcal{C}(\hat{\lambda}_o)$, and there exists an equilibrium with arrival rate $\hat{\lambda}_o$ and mean tip $m = \frac{1}{2\kappa} + U - \hat{C}_o$. In addition, there exists one $\lambda \in (\underline{\lambda}, \mu)$ such that $C(\lambda) = U$ and therefore there exists one more equilibrium with $\lambda \in (\underline{\lambda}, \mu)$ and mean tip $m < \frac{1}{2\kappa}$.
- Otherwise, the cases $U \in [\underline{C}, \hat{C}_o]$ and $U < \underline{C}$ are the same as $U \in [\underline{C}, C(\Lambda_r)]$ and $U < C(\underline{\lambda})$ for $\Lambda_r < \hat{\lambda}_r$, respectively, and their proofs are identical.

Finally, when $\Lambda_r \geq \hat{\Lambda}_r$, then by Lemma 5.12, for every $\lambda \in (\Lambda_r, \mu)$, $U \notin \mathcal{C}(\lambda) = \emptyset$, implying that an equilibrium with $\lambda \in (\Lambda_r, \mu)$ does not exist. □

D.12 Extended explanation for Corollary 5.17

Taking the first two terms of the Taylor expansion of $\Delta(x^2)$ about the point $x \to 0$, we have, for sufficiently small ρ, $\Delta(\rho) \approx \frac{\sqrt{8\rho}}{3}$. Therefore, with the definition of $\hat{\lambda}_r$ and $\hat{\lambda}_o$ (see Proposition 5.9) we have that $\hat{\lambda}_r \approx \frac{9\mu^2}{32\kappa c}$ with $\hat{\lambda}_o \approx \frac{4\sqrt{2\kappa c \Lambda_r^3}}{3\mu}$. Suppose that

$\Lambda_r = \frac{1}{2}(\frac{3\mu^2}{2\sqrt{\kappa c}}(1-\sqrt{\frac{c}{U\mu}}))^{\frac{2}{3}}$. Thus, upon substitution we get $\lim_{\kappa\to\infty}\hat{\lambda}_o = \mu - \sqrt{\frac{c\mu}{U}}$. Note that

$$\hat{C}_o = C(\hat{\lambda}_o) = \frac{c\Lambda_r^2}{(1-\frac{\Lambda_r}{\mu})^2}\frac{1}{\hat{\lambda}_o} + \frac{c}{\mu(1-\frac{\hat{\lambda}_o}{\mu})(1-\frac{\Lambda_r}{\mu})}$$

where the first equality follows by definition and the second follows Equation (D.18), substituting $\lambda = \hat{\lambda}_o$ and $m_\lambda = \frac{1}{2\kappa}$. Thus, when $\kappa \to \infty$, we have $\Lambda_r \to 0$ and $\lim_{\kappa\to\infty}\hat{C}_o = \sqrt{\frac{cU}{\mu}}$. Assuming that $U > \frac{c}{\mu}$ we have for sufficiently small κ, that $U > \hat{C}_o$ and that $\Lambda_r \in (\hat{\lambda}_r, \hat{\Lambda}_r)$ (note that by definition $\hat{\lambda}_o < \hat{\Lambda}_r$ for all $\Lambda_r \in (0,\mu)$ and that $\hat{\lambda}_o$ approaches a constant). Thus, by Proposition 5.14(ii-a), there exists an equilibrium with total arrival rate $\hat{\lambda}_o$ and mean tip $m = \frac{1}{2\kappa} + U - \hat{C}_o$, and for this equilibrium, the tipping wage, $m\hat{\lambda}_o$ satisfies:

$$\lim_{\kappa\to\infty} m\hat{\lambda}_o = \lim_{\kappa\to\infty}(\frac{1}{2\kappa} + U - \hat{C}_o)(\mu - \sqrt{\frac{c\mu}{U}}) = (U - \sqrt{\frac{cU}{\mu}})(\mu - \sqrt{\frac{c\mu}{U}}) = \mu U(1-\sqrt{\frac{c}{\mu U}})^2,$$

which coincides with the socially optimal welfare.

D.13 Derivation of tip-waiting-time correlation

Given an equilibrium tipping distribution $T(t; m, \hat{\lambda}_o)$, with $m = \frac{1}{2\kappa} + U - \hat{C}_o$ (such that one-time customers tip), consider Pearson's correlation coefficient, $\mathrm{Cor}(Z,S)$, where $Z \sim T(t; m, \hat{\lambda}_o)$ is a customer's randomly chosen tip and S is the random waiting (or *sojourn*) time of that customer. Recall, by Lemma 5.12, that when the arrival rate is $\hat{\lambda}_o$, every tip above $1/(2\kappa)$ is rational, and "shifts" the tipping distribution linearly. The correlation coefficient is invariant with respect to shifting, thus for every value κ, it can be assumed, without loss of generality, that $m = \frac{1}{2\kappa}$, and therefore, $Z \sim T(t; \frac{1}{2\kappa}, \hat{\lambda}_o)$. We have

$$\mathrm{Cov}(Z,S) = \mathbb{E}(ZS) - \mathbb{E}(Z)\mathbb{E}(S) \text{ and } \mathrm{Cor}(Z,S) = \frac{\mathbb{E}(ZS) - \mathbb{E}(Z)\mathbb{E}(S)}{\sqrt{\mathrm{Var}(Z)\mathrm{Var}(S)}}. \tag{D.21}$$

Lemma D.1. *Define* $\rho_r = \frac{\Lambda_r}{\mu}$ *and* $\hat{\rho}_o = \frac{\hat{\lambda}_o}{\mu}$, *then:* $\mathbb{E}(S) = \frac{1}{\mu(1-\hat{\rho}_o)}$, $\mathbb{E}(Z) = \frac{1}{2\kappa}$,

$$\mathrm{Var}(Z) = \frac{1}{\kappa^2}(\frac{\kappa c}{\mu}\cdot\frac{\rho_r^2}{\hat{\rho}_o(1-\rho_r)^2} - \frac{1}{4}),\ \mathrm{Var}(S) = \frac{(\rho_r-3)\rho_r^2 + 2\left(\rho_r^2+1\right)\hat{\rho}_o + \left(\rho_r^3 - 3\rho_r^2 + 2\rho_r - 2\right)\hat{\rho}_o^2}{2\mu^2(1-\rho_r)^3\left(1-\hat{\rho}_o\right)^2\hat{\rho}_o}$$

and

$$\mathbb{E}(ZS) = \frac{\rho_r}{\hat{\rho}_o}\frac{1}{\mu}\sqrt{\frac{c}{\kappa\mu}}\frac{2\arcsin(\sqrt{\frac{\rho_r}{2}}) - (1-\rho_r)\sqrt{(2-\rho_r)\rho_r}}{2(1-\rho_r)^2\rho_r}.$$

Proof. We first derive $\mathrm{Var}(S)$. To this purpose, note (from basic Queueing Theory, see Kleinrock (1975)) that $\mathbb{E}(S) = \frac{1}{\mu - \hat{\lambda}_o}$, thus, we have to calculate $\mathbb{E}(S^2)$. From the law of total expectation,

$$\mathbb{E}(S^2) = \frac{\Lambda_r}{\hat{\lambda}_o}\mathbb{E}(S_r^2) + (1 - \frac{\Lambda_r}{\hat{\lambda}_o})\mathbb{E}(S_o^2), \tag{D.22}$$

where S_r and S_o are random variables representing a repeat and a one-time customers' waiting time, respectively.

A repeat customers' waiting time depends on the proportion of customers tipping more than her. Suppose that this customer's tip equals the p-quantile of the tipping distribution, $t_p(\frac{1}{2\kappa}, \hat{\lambda}_o)$ (see Equation (D.9)) and let $S_r(p)$ be her waiting time in the system. The arrival rate of customers tipping more than her is $(1-p)\Lambda_r$. Denote by X_λ the number of customers a moment *after* an arrival to an M/M/1 system with arrival rate λ (and service rate μ), with mean $\overline{x}_\lambda$ and second moment $\overline{x^2}_\lambda$. Then $X_\lambda \sim Geo(1 - \frac{\lambda}{\mu})$. When the customer tipping $t_p(\frac{1}{2\kappa}, \hat{\lambda}_o)$ arrives at the system, her position in the queue is distributed as $X_{(1-p)\Lambda_r}$. In addition, customers who cut her in line continue to flow into the system with rate $(1-p)\Lambda_r$. Denote by B_λ the random variable representing the length of a busy period in an M/M/1 queue with arrival rate λ, and denote its mean by $\overline{b}_\lambda$ and second moment by $\overline{b^2}_\lambda$. The discussed customer will have to wait for a number of $X_{(1-p)\Lambda_r}$ busy periods (each distributed as $B_{(1-p)\Lambda_r}$) until she completes service. Let $B^{(i)}_{(1-p)\Lambda_r}$, $i = 1, 2, ...$ be a sequence of i.i.d random variables distributed as $B_{(1-p)\Lambda_r}$, then

$$\mathbb{E}\big(S_r(p)\big) = \mathbb{E}\left(\sum_{i=1}^{X_{(1-p)\Lambda_r}} B^{(i)}_{(1-p)\Lambda_r}\right) = \overline{x}_{(1-p)\Lambda_r}\overline{b}_{(1-p)\Lambda_r},$$

and

$$\begin{aligned}\mathbb{E}\left(S_r(p)^2\right) &= \mathbb{E}\left(\left(\sum_{i=1}^{X_{(1-p)\Lambda_r}} B^{(i)}_{(1-p)\Lambda_r}\right)^2\right) = \mathbb{E}(X_{(1-p)\Lambda_r})\overline{b^2}_{(1-p)\Lambda_r} + 2\mathbb{E}\left(\sum_{i<j}^{X_{(1-p)\Lambda_r}} B^{(i)}_{(1-p)\Lambda_r}B^{(j)}_{(1-p)\Lambda_r}\right) \\ &= \overline{x}_{(1-p)\Lambda_r}\overline{b^2}_{(1-p)\Lambda_r} + \mathbb{E}\big(X_{(1-p)\Lambda_r}(X_{(1-p)\Lambda_r} - 1)\big)(\overline{b}_{(1-p)\Lambda_r})^2 \\ &= \overline{x}_{(1-p)\Lambda_r}\overline{b^2}_{(1-p)\Lambda_r} + (\overline{x^2}_{(1-p)\Lambda_r} - \overline{x}_{(1-p)\Lambda_r})(\overline{b}_{(1-p)\Lambda_r})^2\end{aligned} \tag{D.23}$$

From Queueing Theory (see Haviv (2013)) we have

$$\overline{x}_\lambda = \frac{1}{1-\frac{\lambda}{\mu}}, \qquad \overline{x^2}_\lambda = \frac{1}{(1-\frac{\lambda}{\mu})^2}, \qquad \overline{b}_\lambda = \frac{1/\mu}{1-\frac{\lambda}{\mu}}, \qquad \overline{b^2}_\lambda = \frac{2/\mu^2}{(1-\frac{\lambda}{\mu})^3},$$

and upon substitution we get

$$\mathbb{E}\big(S_r(p)\big) = \frac{1}{\mu(1-(1-p)\frac{\Lambda_r}{\mu})^2}, \text{ and } \mathbb{E}\big(S_r(p)^2\big) = \frac{(1-p)\frac{\Lambda_r}{\mu} + 2}{\mu^2(1-(1-p)\frac{\Lambda_r}{\mu})^4}. \tag{D.24}$$

Integrating over $p \in [0,1]$ (by the law of total expectation) gives the second moment of the repeat customers' waiting time:

$$\mathbb{E}(S_r^2) = \int_0^1 \mathbb{E}\left(S_r(p)^2\right) dp = \int_0^1 \frac{(1-p)\frac{\Lambda_r}{\mu}+2}{\mu^2(1-(1-p)\frac{\Lambda_r}{\mu})^4} dp = \frac{4-(3-\frac{\Lambda_r}{\mu})\frac{\Lambda_r}{\mu}}{2\mu^2(1-\frac{\Lambda_r}{\mu})^3}. \tag{D.25}$$

Consider now a one-time customer arriving at the system (recall that the overall rate of arrival is $\hat{\lambda}_o$). The number of customers in the system after her arrival is distributed as $X_{\hat{\lambda}_o}$. Repeat customers continue to arrive with rate Λ_r. Thus, this customer has to wait a time length equally distributed as $X_{\hat{\lambda}_o}$ busy periods of a queue with arrival rate Λ_r. As in Equation (D.23), we obtain

$$\mathbb{E}\left(S_o^2\right) = \mathbb{E}\left(\left(\sum_{i=1}^{X_{\hat{\lambda}_o}} B_{\Lambda_r}^{(i)}\right)^2\right) = \overline{x}_{\hat{\lambda}_o}\overline{b^2}_{\Lambda_r} + (\overline{x^2}_{\hat{\lambda}_o} - \overline{x}_{\hat{\lambda}_o})(\overline{b}_{\Lambda_r})^2 = \frac{2-(\frac{\Lambda_r}{\mu}+1)\frac{\hat{\lambda}_o}{\mu}}{\mu^2(1-\frac{\Lambda_r}{\mu})^3\left(1-\frac{\hat{\lambda}_o}{\mu}\right)^2}. \tag{D.26}$$

Substituting (D.25) and (D.26) in Equations (D.22) and rearranging, we arrive at

$$\begin{aligned}\mathbb{E}\left(S^2\right) &= \frac{\Lambda_r}{\hat{\lambda}_o}\cdot\frac{4-(3-\frac{\Lambda_r}{\mu})\frac{\Lambda_r}{\mu}}{2\mu^2(1-\frac{\Lambda_r}{\mu})^3} + (1-\frac{\Lambda_r}{\hat{\lambda}_o})\frac{2-(\frac{\Lambda_r}{\mu}+1)\frac{\hat{\lambda}_o}{\mu}}{\mu^2(1-\frac{\Lambda_r}{\mu})^3(1-\frac{\hat{\lambda}_o}{\mu})^2} \\ &= \frac{\rho_r\left(\rho_r^2-3\rho+4\right)\left(\hat{\rho}_o-1\right)^2 + 2\left(\rho_r-\hat{\rho}_o\right)\left((\rho_r+1)\hat{\rho}_o-2\right)}{2\mu^2(1-\rho_r)^3\left(1-\hat{\rho}_o\right)^2\hat{\rho}_o},\end{aligned}$$

where $\rho_r = \frac{\Lambda_r}{\mu}$ and $\hat{\rho}_o = \frac{\hat{\lambda}_o}{\mu}$. Thus, with $\mathbb{E}(S) = \frac{1}{\mu(1-\hat{\rho}_o)}$ we get

$$\mathrm{Var}(S) = \mathbb{E}\left(S^2\right) - \mathbb{E}(S)^2 = \frac{(\rho_r-3)\rho_r^2 + 2(\rho_r^2+1)\hat{\rho}_o + (\rho_r^3-3\rho_r^2+2\rho_r-2)\hat{\rho}_o^2}{2\mu^2(1-\rho_r)^3(1-\hat{\rho}_o)^2\hat{\rho}_o}. \tag{D.27}$$

Since $Z \sim T(t;\frac{1}{2\kappa},\hat{\lambda}_o)$ in equilibrium satisfies rational tipping (see Section 5.3.3), we have that $\mathbb{E}(Z) = \frac{1}{2\kappa}$, and

$$\mathrm{Var}(Z) = \mathbb{E}\left(Z^2\right) - \mathbb{E}(Z)^2 = ET^2(\frac{1}{2\kappa},\hat{\lambda}_o) - (\frac{1}{2\kappa})^2 = \frac{c\Lambda_r^2}{\kappa\hat{\lambda}_o(\mu-\Lambda_r)^2} - \frac{1}{4\kappa^2} = \frac{1}{\kappa^2}(\frac{\kappa c}{\mu}\cdot\frac{\rho_r^2}{\hat{\rho}_o(1-\rho_r)^2} - \frac{1}{4}) \tag{D.28}$$

where the third equality follows Equation (D.17). Last, we calculate $\mathbb{E}(ZS)$. Recall the definition of $t_p(\frac{1}{2\kappa},\hat{\lambda}_o)$ from Equation (D.9) and the expression for $\mathbb{E}\big(S_r(p)\big)$ from

Equation (D.24). Then,

$$\begin{aligned}\mathbb{E}(ZS) &= \frac{\Lambda_r}{\hat{\lambda}_o}\mathbb{E}\big(ZS \mid Z \in (0,\overline{t}]\big) + (1-\frac{\Lambda_r}{\hat{\lambda}_o})\mathbb{E}(ZS \mid Z=0) = \frac{\Lambda_r}{\hat{\lambda}_o}\int_{p=0}^{1} t_p(\frac{1}{2\kappa},\Lambda_r)\cdot\mathbb{E}\big(S_r(p)\big)dp \\ &= \frac{\Lambda_r}{\hat{\lambda}_o}\int_{p=0}^{1}\frac{1}{\mu(1-(1-p)\rho_r)^2}\sqrt{\frac{c}{\kappa\mu}\left(\frac{1}{(1-\rho_r)^2}-\frac{1}{(1-(1-p)\rho_r)^2}\right)}dp \\ &= \frac{\rho_r}{\hat{\rho}_o}\cdot\frac{1}{\mu}\cdot\sqrt{\frac{c}{\kappa\mu}}\frac{2\arcsin\left(\sqrt{\frac{\rho_r}{2}}\right)-(1-\rho_r)\sqrt{(2-\rho_r)\,\rho_r}}{2\,(1-\rho_r)^2\,\rho_r}.\end{aligned} \tag{D.29}$$

We further note that $\hat{\lambda}_o = 2\Lambda_r\sqrt{\frac{\kappa c}{\mu}}\Delta(\frac{\Lambda_r}{\mu})$, thus, $\hat{\rho}_o = 2\rho_r\sqrt{\frac{\kappa c}{\mu}}\Delta(\rho_r)$. Assuming that $T(t;\frac{1}{2\kappa},\hat{\lambda}_o)$ is an equilibrium distribution, it must therefore hold that $\kappa \in [(2\sqrt{c/\mu}\Delta(\rho_r))^{-2}, (\frac{2}{\mu}\rho_r\sqrt{c/\mu}\Delta$ otherwise either $\hat{\lambda}_o > \mu$ or $\Lambda_r < \hat{\lambda}_r$, contrary to Proposition 5.14(ii-a). □

Numerical illustration: We illustrate $\sqrt{\mathrm{Var}(Z)}$ and $\sqrt{\mathrm{Var}(S)}$, $\mathrm{Cov}(Z,S)$, $\mathrm{Cor}(Z,S)$, for $c=1$, $\mu=1$ and $\Lambda_r = 1/4$:

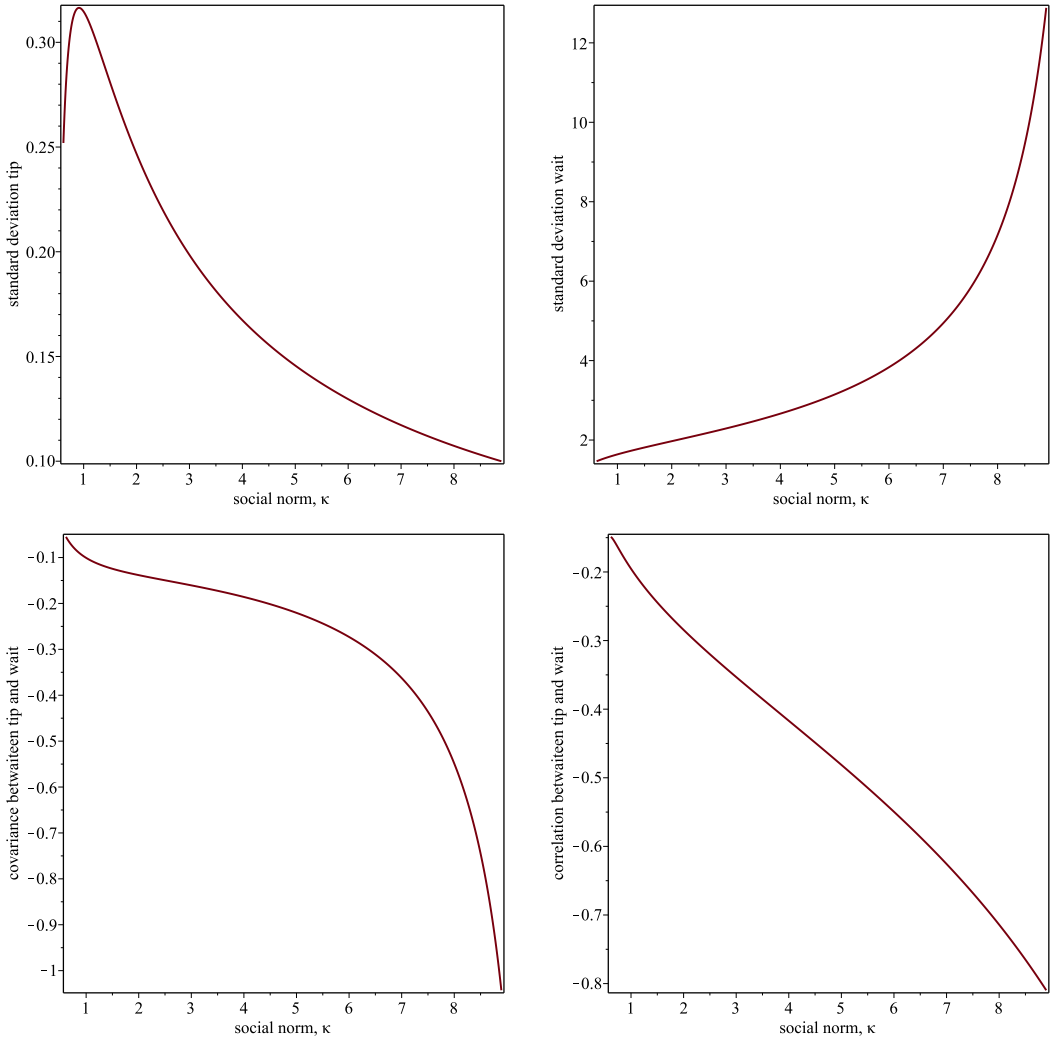


FIGURE D.1: Standard deviation of tips, waiting time, covariance and correlation between tips and waiting time for $c = 1$, $\mu = 1$ and $\Lambda_r = 1/4$ as a function of the strength of the social norm; κ, for the equilibrium in which one-time customers join and tip; the arrival rate is $\hat{\lambda}_o$ and tip is $1/(2\kappa) + U - \hat{C}_o$.

Bibliography

Refael Hassin and Moshe Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, 2003.

Refael Hassin. *Rational Queueing*. CRC Press, Boca Raton, 2016.

Refael Hassin and Ran I Snitkovsky. Social and monopoly optimization in observable queues. *Operations Research, to appear*, 2019.

Ilan Adler and Pinhas Naor. Social optimization versus self-optimization in waiting lines. *Technical Report*, 1969.

Olga Boudali and Antonis Economou. Optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes. *European Journal of Operational Research*, 218:708–715, 2012.

John Hasenbein and Ying Chen. Parameter uncertainty in Naor's model. *working paper*, 2016.

Husheng Li and Zhu Han. Socially optimal queuing control in cognitive radio networks subject to service interruptions: To queue or not to queue. *IEEE Transactions on Wireless Communications*, 10:1656–1666, 2011.

Wei Sun and Shiyong Li. Customer threshold strategies in observable queues with partial information of service time. *Information Computing and Applications*, 307:456–462, 2012.

Wei Sun, Yan Wang, Kaibo Yuan, and Shiyong Li. Customer joining-balking strategies in an observable queue with partial service time information. *International Journal of Applied Mathematics*, 48, 2018.

Jinting Wang, Zhe George Zhang, and Zhengwu Zhang. Performance analysis of a queue with strategic customers under quadratic utility criterion. *working paper*, 2014.

Zhengwu Zhang, Jinting Wang, and Feng Zhang. Equilibrium customer strategies in the single-server constant retrial queue with breakdowns and repairs. *Mathematical Problems in Engineering*, 2014:14 pages, 2014.

Bernardo D'Auria and Spyridoula Kanta. Pure threshold strategies for a two-node tandem network under partial information. *Operations Research Letters*, 43:467–470, 2015.

Yoav Kerner. Equilibrium joining probabilities for an M/G/1 queue. *Games and Economic Behavior*, 71:521–526, 2011.

Bara Kim and Jeongsim Kim. Equilibrium strategies for a tandem network under partial information. *Operations Research Letters*, 44:532–534, 2016.

Antonis Economou and Spyridoula Kanta. Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. *Queueing Systems*, 59:237–269, 2008a.

Antonis Economou and Spyridoula Kanta. Equilibrium customer strategies and social-profit maximization in the single-server constant retrial queue. *Naval Research Logistics*, 58:107–122, 2011.

Niels C. Knudsen. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica*, 40:515–528, 1972.

Pinhas Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

András Simonovits. Self- and social optimization in queues. *Studia Scientiarum Mathematicarum Hungarica*, 11:131–138, 1976.

Noel M. Edelson and David K. Hildebrand. Congestion tolls for Poisson queueing processes. *Econometrica*, 43:81–92, 1975.

Refael Hassin and Ran I Snitkovsky. Strategic customer behavior in a queueing system with a loss subsystem. *Queueing Systems*, 86(3-4):361–387, 2017.

Laurens Debo and Ran I Snitkovsky. Tipping in service systems: The role of a social norm. *Available at SSRN 3287862*, 2018.

Garrett Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

Søren G. Johansen and Shaler Stidham Jr. Control of arrivals to a stochastic input-output system. *Advances in Applied Probability*, 12:972–999, 1980.

Shaler Stidham Jr. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30:705–713, 1985.

Eitan Altman and Refael Hassin. Non-threshold equilibrium for customers joining an M/G/1 queue. *in Proceedings of 10th International Symposium on Dynamic Game and Applications*, 2002.

Steven A. Lippman and Shaler Stidham Jr. Individual versus social optimization in exponential congestion systems. *Operations Research*, 25:233–247, 1977.

Shaler Stidham Jr. Socially and individually optimal control of arrivals to a GI/M/1 queue. *Management Science*, 24:1598–1610, 1978.

Haim Mendelson and Uri Yechiali. Controlling the GI/M/1 queue by conditional acceptance of customers. *European Journal of Operational Research*, 7:77–85, 1981.

Uri Yechiali. On optimal balking rules and toll charges in the GI/M/1 queue. *Operations Research*, 19:349–370, 1971.

Uri Yechiali. Customers' optimal joining rules for the GI/M/s queue. *Management Science*, 18:434–443, 1972.

Susan H. Xu and George J. Shanthikumar. Optimal expulsion control: a dual approach to admission control of an ordered-entry system. *Operations Research*, 41:1137–1152, 1993.

Chia-Li Wang. On socially optimal queue length. *Management Science*, 62:899–903, 2016.

Refael Hassin. On the optimality of first-come last-served queues. *Econometrica*, 53: 201–202, 1985.

Moshe Haviv and Binyamin Oz. Regulating an observable M/M/1 queue. *Operations Research Letters*, 44:196–198, 2016.

Jung Hyun Kim, Hyun-Soo Ahn, and Rhonda Righter. Managing queues with heterogeneous servers. *Journal of Applied Probability*, 48:435–452, 2011.

Arthur De Vany. Uncertainty, waiting time, and capacity utilization: A stochastic theory of product quality. *Journal of Political Economy*, 84:523–542, 1976.

Hong Chen and Murray Frank. Monopoly pricing when customers queue. *IIE Transactions*, 36:569–581, 2004.

Moshe Shaked and George J. Shanthikumar. *Stochastic Orders*. Springer, New York, 2007.

Eitan Altman and Nahum Shimkin. Individual equilibrium and learning in processor sharing systems. *Operations Research*, 46:776 – 784, 1998.

Apostolos Burnetas and Yiannis Dimitrakopoulos. Strategic equilibria in queues with dynamic service rate and full information. *working paper*, 2018.

Moshe Haviv. *Queues – A course in queueing theory.* Springer, New York, 2013.

Mark B. Garman. Market microstructure. *Journal of Financial Economics*, 3:257–275, 1976.

Antonis Economou and Spyridoula Kanta. Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Operations Research Letters*, 36:696–699, 2008b.

Onno J. Boxma. Joint distribution of sojourn time and queue length in the M/G/1 queue with (in)finite capacity. *European Journal of Operational Research*, 16:246–256, 1984.

Yoav Kerner. The conditional distribution of the residual service time in the M_n/G/1 queue. *Stochastic Models*, 24:364–375, 2008.

Jiaming Xu and Bruce Hajek. The supermarket game. *Stochastic Systems*, 3:405–441, 2013.

Ricky Roet-Green and Refael Hassin. The impact of inspection cost on equilibrium, revenue, and social welfare in a single server queue. *Operations Research*, 2014.

Stephen C. Littlechild. Optimal arrival rate in a simple queueing system. *International Journal of Production Research*, 12,:391–397, 1974.

Chuong T. Do, Nguyen H. Tran, Mui Van Nguyen, Choong Seon Hong, and Sungwon Lee. Social optimization strategy in unobserved queueing systems in cognitive radio networks. *IEEE Communications Letters*, 16:1944–1947, 2012.

Oussama Habachi and Yezekael Hayel. Optimal opportunistic sensing in cognitive radio networks. *IET Communications*, 6:797–804, 2012.

Yezekael Hayel, Dominique Quadri, Tania Jimnez, and Luce Brotcorne. Decentralized optimization of last-mile delivery services with non-cooperative bounded rational customers. *Annals of Operations Research*, 239:451469, 2014.

Krishna Jagannathan, Ishai Menache, Eytan Modiano, and Gil Zussman. Non-cooperative spectrum access - the dedicated vs. free spectrum choice. *IEEE Journal on Selected Areas in Communications*, 30:2251–2261, 2012.

Simon Haykin. Cognitive radio: Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23:201–220, 2005.

Uri Yechiali and Pinhas Naor. Queuing problems with heterogeneous arrivals and service. *Operations Research*, 19:722–734, 1971.

Michael Lynn. Tipping in restaurants and around the globe: An interdisciplinary review. In Morris Altman, editor, *Handbook of Contemporary Behavioral Economics: Foundations and Developments*, chapter 31, pages 626–643. M.E.Sharpe, New York, 2006.

Nancy Star. *The International Guide to Tipping.* Berkley Books, New York, 1988.

Ofer H. Azar and Yossi Tobol. Tipping as a strategic investment in service quality: An optimal-control analysis of repeated interactions in the service industry. *Southern Economic Journal*, 75:246–260, 2008.

Maggie R. Jones. Measuring the effects of the tipped minimum wage using w-2 data. *CARRA Working Paper Series*, 2016-03, 2016.

Ofer H. Azar. The history of tipping – from sixteenth-century England to United States in the 1910s. *The Journal of Socio-Economics*, 33:745–764, 2004a.

Ofer H. Azar. Why pay extra? tipping and the importance of social norms and feelings in economic theory. *The Journal of Socio-Economics*, 85:141–173, 2005a.

Heidi Shierholz, David Cooper, Julia Wolfe, and Ben Zipperer. Employers would pocket $5.8 billion of workers tips under trump administrations proposed tip stealing rule. *Economic Policy Institute Report*, December 12, 2017.

Ana Wood. A fairer system for tipping. as long as american diners are paying gratuity, this is how it should be done. *The Atlantic*, January 15, 2017.

Maria Yagoda. Everything you need to know about the restaurant tipping debate. *Food and Wine*, May 01,, 2018.

Orn B. Bodvarsson and William A. Gibson. Economics and restaurant gratuities: Determining tip rates. *The American Journal of Economics and Sociology*, 56:187–203, 1997.

Michael Lynn and Michael McCall. Gratitude and gratuity: a meta-analysis of research on the service-tipping relationship. *Journal of Socio-Economics*, 29:203–214, 2000.

John H. Pencavel. Work effort, on-the-job screening, and alternative methods of remuneration. In *35th Anniversary Retrospective*, pages 537–570. Emerald Group Publishing Limited, 2015.

Nancy Jacob and Alfred Page. Production, information costs, and economic organization: The buyer monitoring case. *American Economic Review*, 70:476–478, 1980.

Andrew Schotter. *Microeconomics: A Modern Approach.* Addison-Wesley, Reading, MA, 2000.

Z. Schwartz. The economics of tipping: Tips, profits, and the market's demand-supply equilibrium. *Tourism Economy*, 3:265–279, 1997.

Michael Lynn and Shuo Wang. The indirect effects of tipping policies on patronage intentions through perceived expensiveness, fairness, and quality. *Journal of Economic Psychology*, 39:62–71, 2013.

David E. Sisk and Edward C. Gallick. Tips and commissions: A study in economic contracting. Bureau of Economics, Federal Trade Commission, 1985.

Uri Ben-Zion and Edi Karni. Tip payments and the quality of service. *Essays in labor market analysis*, The Froeder Institute for Economic Research. Tel Aviv University, Tel Aviv:37–44, 1977.

Ofer H. Azar. The social norm of tipping: A review. *Journal of Applied Social Psychology*, 37:380–402, 2007.

Ofer H. Azar. What sustains social norms and how they evolve? The case of tipping. *Journal of Economic Behavior & Organization*, 54:49–64, 2004b.

Ofer H. Azar. Strategic behavior and social norms in tipped service industries. *The B.E. Journal of Economic Analysis & Policy*, 8:Article 7, 2008.

Michael Conlin, Michael Lynn, and Ted O'Donoghue. The norm of restaurant tipping. *Journal of Economic Behavior & Organization*, 52:297–321, 2003.

Joanne M. May. Tip or treat: A study of factors affecting tipping behavior. Master's thesis, Loyola University, Chicago, 1978.

Michael Lynn and Andrea Grassman. Restaurant tipping: An examination of three 'rational' explanations. *Journal of Economic Psychology*, 11:169–181, 1990.

Orn B. Bodvarsson and William A. Gibson. Gratuities and customer appraisal of service: Evidence from minnesota restaurants. *Journal of Socio-Economics*, 23:287–303, 1994.

Clive Seligman, Joan E. Finegan, J. Douglas Hazlewood, and Mark Wilkinson. Manipulating attributions for profit: A field test of the effects of attributions on behavior. *Social Cognition*, 3:313–321, 1986.

Francis T. Lui. An equilibrium queueing model of bribery. *Journal of Political Economy*, 93:760–781, 1985.

Amihai Glazer and Refael Hassin. Stable priority purchasing in queues. *Operations Research Letters*, 4:285–288, 1986.

Tingliang Huang and Ying-Ju Chen. Service systems with experience-based anecdotal reasoning consumers. *Production and Operations Management*, 25(5):778–790, 2014.

Liu Yang, Pengfei Guo, and Yulan Wang. Service pricing with loss-averse customers. *Operations Research*, 66, 2018.

Michael Lynn, Patrick Jabbour, and Woo Gon Kim. Who uses tips as a reward for service and when? an examination of potential moderators of the service-tipping relationship. *Journal of Economic Psychology*, 33:90–103, 2012.

Philipp Afèche, Opher Baron, Joseph Milner, and Ricky Roet-Green. Pricing and prioritizing time-sensitive customers with heterogeneous demand rates. Working paper, 2015.

B.J. Ruffle. Gift giving with emotions. *Journal of Economic Behavior and Organization*, 39:399–420, 1999.

Leonard Kleinrock. *Queueing Systems, Volume 2: Computer Applications*. John Wiley & Sons Inc., New York, 1976.

Ofer H. Azar. The social norm of tipping: Does it improve social welfare? *Journal of Economics*, 85:141–173, 2005b.

John F.C. Kingman. The effect of queue discipline on waiting time variance. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 58, pages 163–164. Cambridge University Press, 1962.

Jacob W. Cohen. *The Single Server Queue*. North-Holland Publishing Company, Amsterdam, 1982.

Moshe Haviv and Ya'acov Ritov. Externalities, tangible externalities, and queue disciplines. *Management Science*, 44:850–858, 1998.

Leonard Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons Inc., Canada, 1975.