UNIVERSITY OF CALIFORNIA,
IRVINE

Interdisciplinary Research in Operations Management: Applications in Healthcare, Retailing and On-demand Service Platforms

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Management

by

Jiaru Bai

Dissertation Committee:
Professor L. Robin Keller, Co-Chair
Professor Shuya Yin, Co-Chair
Professor Kut (Rick) C. So

2017

DEDICATION

To my parents for their continuing support and love. They allowed me to grow and thrive.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

Looking back on the past 5 years of my Ph.D. studies, I am very grateful and would like to acknowledge and thank all those people who have encouraged, mentored and helped me.

First of all, I would like to extend my deepest gratitude to my advisors, Professor L. Robin Keller and Professor Shuya Yin. Both of them are my committee co-chairs. Professor Keller has been very supportive and altruistic and is always there when I need help. She is one of the most influential people in my life as both my life coach and academic advisor. Professor Keller often encourages me to follow my own interests and she is very open minded. I always can share my thoughts with her without reservation. Professor Keller is absolutely a role model for me to follow in how to deal with other people.

Second, I would also like to express my sincerest appreciation to Professor Yin, who has sparked my interest in retailing. Professor Yin has provided me with invaluable support and encouragement during my Ph.D. studies. Also, Professor Yin has taught me valuable knowledge and given me tremendous help in my job search. I am most inspired by her enthusiasm for both research and teaching. Professor Yin has constantly steered me in the right direction whenever I encountered a problem. Without her guidance, this dissertation would not have been possible.

Third, I am indebted to my committee member, Professor Rick So, who has provided my with excellent guidance and tremendous help in my research. Professor So is very knowledgeable and extremely supportive in every meeting. Professor So is always responsive and efficient. Doing research with him is always enjoyable and indeed it is one of my most memorable experiences.

I am also thankful to my co-authors, Professors Christopher S. Tang, Haresh Gurnani, Cristina del Campo and many others for the guidance and assistance on my research projects. Special thanks to Professors Carlton Scott, John Turner and Luyi Gui, who have taught me useful knowledge and provided me with helpful suggestions.

Finally, I am grateful to Didi Chuxing (www.xiaojukeji.com) for providing us some sample data.

CURRICULUM VITAE

Jiaru Bai

EDUCATION

Ph.D. in Management **2017**
Paul Merage School of Business, University of California, Irvine *Irvine, California*

M.S. in Statistics **2014**
Donald Bren School of Information & Computer Sciences, UC Irvine *Irvine, California*

B.S. in Engineering **2012**
Honors College, Beihang University *Beijing, China*

B.S. in Economics **2012**
National School of Development, Peking University *Beijing, China*

RESEARCH INTEREST

Business Analytics
Service Operations Management
Medical Decision Analysis
Behavioral Operations Management
Operations Management and Marketing Interfaces
Multi-channel Retailing

AWARDS AND HONORS

First Prize, Student Paper Competition of the College of SCM at the POMS Conference	2017
Doctoral Student Fellowship, UC Irvine	2012-2016
Ray Watson Fellowship, UC Irvine	2016

MEMBERSHIP

Institute of Operations Research and the Management Sciences
Manufacturing and Service Operations Management Society
Production and Operations Management Society
Decision Sciences Institute

PROFESSIONAL ACTIVITY

Associated Certified Analytics Professional (aCAP)	2016-2017
Session Chair, 2017 POMS Conference	2017
Session Chair, 2017 INFORMS Computing Society's Annual Meeting	2017

ABSTRACT OF THE DISSERTATION

Interdisciplinary Research in Operations Management: Applications in Healthcare, Retailing and On-demand Service Platforms

By

Jiaru Bai

Doctor of Philosophy in Management

University of California, Irvine, 2017

Professor L. Robin Keller, Co-Chair
Professor Shuya Yin, Co-Chair

This dissertation consists of three essays on applications of interdisciplinary research in operations management. The first essay addresses issues in healthcare. Our goal is to evaluate the cost-effectiveness of bevacizumab compared to the baseline treatment with only chemotherapy in recurrent/persistent and metastatic cervical cancer using recently reported updated survival and toxicology data. We developed a Markov model with 5 patient health states for both treatments. With data based on the Gynecologic Oncology Group 240 randomized trials and the 2013 MediCare Services Drug Payment Table and Physician Fee Schedule, we present monthly transition probabilities and cost data. Our results show that chemotherapy plus bevacizumab can delay progression, but incur more complications.

The second essay lies at the interface between operations management and marketing. We aim to understand the tradeoffs in offering outlet stores. In particular, we study how much differentiation should be kept between the main and outlet stores from three perspectives: price, product and location. We find that an outlet store is more likely to be opened when travel sensitivity is lower or costs associated with it are lower. Moreover, offering an outlet store encourages the firm to improve the quality of the product sold in the main store as

to reduce the cannibalization effect. We also observe that location differentiation has a substitution effect on quality and price differentiation.

In the third essay, we study several operational challenges for the on-demand service platforms. We consider a situation when an on-demand service platform uses earning sensitive independent providers with heterogeneous reservation prices to serve its time and price sensitive customers with heterogeneous valuation of the service. We present a queueing model with endogenous supply and endogenous demand to model this on-demand service platform. Based on our analysis, we find that it is optimal for the platform to charge a higher price, pay a higher wage, and offer a higher payout ratio when the potential customer demand increases. We use a set of actual data from a large on-demand ride-hailing platform in numerical experiments to illustrate some of our main insights.

Chapter 1

A Markov Model to Evaluate Cost-Effectiveness of Bevacizumab in Advanced Cervical Cancer

1.1. Introduction

Women with recurrent and metastatic cervical cancer have an extremely poor prognosis and comprise a population for whom effective therapy has remained a high unmet clinical need. In 2009, Gynecologic Oncology Group (GOG) 204 established cisplatin in combination with paclitaxel as the chemotherapy standard [1]. Although responses rates (RR) of up to 36% can be achieved in platinum-nave patients, for the most part they are not durable, with early progression, rapid deterioration of quality of life (QoL), and death within 7 to 12 months being the rule.2 Furthermore, due to acquired drug resistance associated with prior platinum exposure during cisplatin-based chemoradiation for locally advanced disease, re-treatment with platinum-based therapy at recurrence has been shown to be less effective [2].

In an effort to harness the therapeutic potential of targeting the vascular endothelial growth factor (VEGF) pathway to inhibit tumor-associated angiogenesis, GOG protocol 240 was activated in 2009 throughout the United States and Spain [2]. The primary endpoints were overall survival (OS) and toxicity. In early 2013 it was reported that the arms administering the anti-VEGF humanized monoclonal antibody, bevacizumab were associated with a statistically significant improvement in OS (17 vs 13.3 mos; hazard ratio (HR) of death 0.71 (98% CI, 0.54-0.95;1-sided p=0.004), PFS (8.2 vs 5.9 mos; HR of progression 0.67 (95 % CI, 0.54-0.82); 2-sided p=0.002), and RR (48 % vs 36 %; relative probability of response 1.35; (95 % CI, 1.08-1.68; 2-sided p=0.008), without any significant deterioration in QoL.2,3 The major treatment-related toxicities included fistula (8.6 %), thromboembolism (8.2 %) and manageable hypertension (25 %) [2].

On August 18, 2014, the U.S. FDA approved to expand the label of bevacizmab to include cervical cancer [4]. However, the potential for serious adverse events, including intestinal perforation, fistula, delayed wound healing, hemorrhage, hypertension, proteinuria, and thromboembolism, remain of considerable concern. We decided to study the cost-effectiveness of bevacizumab in advanced cervical cancer.

1.2. Methods

A Markov decision tree using the TreeAge Pro program was created to perform a cost-effectiveness analysis of chemotherapy versus chemotherapy plus bevacizumab for first treatment of recurrent/persistent or metastatic cervical cancer using the data from the GOG 240 study [3,8,9]. Costs were obtained from the Center for MediCare Services Drug Payment Table and Physician Fee Schedule. Only 2013 direct costs were used; billed charges and indirect costs were not featured (Table 1.1).

Table 1.1: Cost for cancer therapy and management of complications

HEALTH STATES	CHEMORx ONLY		CHEMORx+BEVACIZUMAB	
Cancer Therapy*	$524		$7,540	
Treatment of Hypertension**	$285		$285	
Weighted Thromboembolism***	$4,261 x 4/6		$4,261 x 18/37	
Weighted Fistula****	$16,000/3 x 2/6		$16,000/3 x 19/37	
	Total Cost per 28-Day Cycle	Cost Breakdown	Total Cost per 28-Day Cycle	Cost Breakdown
Respond	$524		$7,540	
Progress	$262		$262	
Limited complications	809	ChemoRx + Treatable hypertension	$7,825	ChemoRx + Treatable hypertension
Severe complications	$4,157	Weighted costs	$4,331	Weighted costs
Die	$0	-	$0	-

*cost of chemotherapy alone or chemotherapy plus bevacizumab

**cost of anti-hypertensive medication

***cost of hospitalization, imaging studies, and anti-coagulation; weighted estimation based on analysis of adverse events from the primary manuscript

****cost of imaging studies, colostomy and 3 days of hospitalization; weighted estimation based on analysis of adverse events from the primary manuscript

Note: in the GOG 240 population, approximately 1 of every 3 patients who developed GI-vaginal fistula underwent fecal diversion via colostomy

The GOG 240 study was used to maintain homogeneity of the target population, (i.e., untreated patients with advanced cervical cancer). The model was designed from the perspective of the patient and the health service payer. Our time horizon was 60 months (i.e., 5 years).

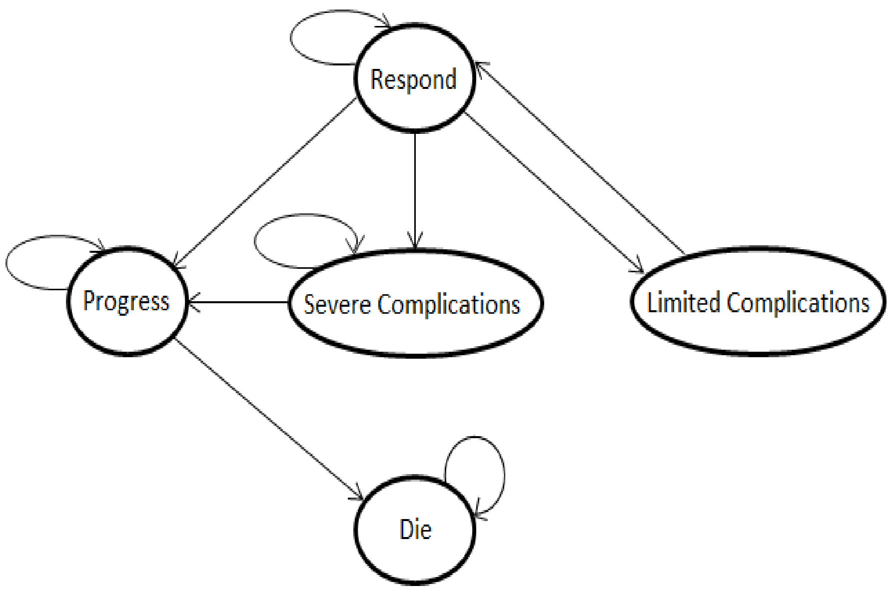


Figure 1.1: Markov diagram for women with advanced cervical cancer treated on Gynecologic Oncology Group protocol 240. Circular arrows indicate that patients can stay in that state with some probability for more than one cycle. Our model has the feature that a patient can stay in any of the 5 states for more than 1 cycle. Die (or death) is an absorbing state, which means that once a patient enters that state she will never leave that state. As time passes, most of the patients will go to the die state.

In the Markov model, five possible health states exist: respond, progress, limited complications, severe complications, and die (Figure **??**). A patient is modeled as being in one state during a month, and the patient may transition to a different state with some probability in the next month. Patients who respond to treatment may remain in response or experience complications (limited or severe) or progress in the next cycle. Those who progress are removed from clinical trial participation and may possibly receive salvage/palliative therapy but ultimately die. Limited complications include hypertension for which they receive pharmacologic management. Because in GOG 240 no patients were taken off study for treatment-induced hypertension, those who develop limited complications in our model recover and may continue to respond or progress. Severe complications are represented by thromboembolism and fistula. Patients with severe complications end their clinical trial participation and receive pharmacologic and/or surgical management of their complication. Finally, we assume that the only way for a patient to go to the die state is following progression and therefore we did not factor in death from other causes. Importantly, the number of treatment-related deaths in the chemotherapy and chemotherapy plus bevacizumab arms in GOG 240 were equal. The

Table 1.2: Assumptions made when developing the Markov model

1	Inclusion of all study subjects (n=452), including those who dropped out
2	Patients assessed for progression every month by physical examination and every 2 months by imaging (RECIST)
3	Length of a cycle in the Markov model is 1 month
4	Patients can only get one complication at a time
5	The only way to go to the complications states is from respond
6	Once a patient progresses there is no way for her to go back to respond
7	Existence of a severe complication excludes listing of a limited complication if one occurs simultaneously
8	If the patient develops limited complications she can go to respond in one month
9	Can only reach the die state from the progress state
10	All transition probabilities are stable and do not change month to month
11	Although imaging studies may be used in patients who develop fistula (eg., rectovaginal, vesicovaginal, etc), these studies are not required to make the diagnosis and therefore the costs of imaging in this setting are not included in the model

model was run over 5 years (60 months). A patient starts in the respond state, then each month either stays in the same state or moves to a new one. Each month the cost of treatment is incurred and a helath utility level is experienced. After 60 months, the total cost is calcultated and the total months lived as well as the equivalent quality adjusted months are added up. The results are the expected costs and months, averaged over all patients.

To ensure validity of our Markov model, several assumptions were made (Table 1.2). Patients who respond cannot directly go to the death state without first passing through the progression state. Because the protocol-specified treatment occurred at 21-day intervals in GOG 240, we rounded up and made the length of a cycle in the Markov model one month. Based on the primary GOG 240 manuscript, the mean number of cycles for patients receiving chemotherapy alone is 6, and for those receiving chemotherapy plus bevacizumab the mean number of cycles is 7. Therefore, the costs associated with managing complications included medication for blood pressure control (grade 2 or higher hypertension occurred in 25% treated with bevacizumab versus 1.8% receiving chemotherapy alone), imaging and anti-coagulation for thromboembolism (grade 3 or higher thromboembolism occurred in 8.2% receiving bevacizumab versus 1.8% treated with chemotherapy alone), and colostomy for some patients with fistula (grade 2 or higher fistula occurred in 8.6% treated with bevacizumab versus 1% receiving chemotherapy alone [6]). The incidence of febrile neutropenia and treatment-related deaths did not differ between the chemotherapy alone and chemotherapy plus bevacizumab cohorts in GOG 240.

As stated above, in our model to match the cycle of cost/treatment (3 weeks plus several days of potential delays (eg., neutropenia requiring repeat blood draws, etc), we set the cycle to be roughly one month. Based on the 1-month/cycle of therapy methodology, using the data reported in GOG protocol 240, the transition probabilities listed in Table 1.3 were obtained.

1.2.1 Health Utilities

In the Markov model, the patient experiences a health "reward or "utility in each month, representing the effectiveness of the treatment which depends on the health state during that month. The patients overall effectiveness is the sum of these utilities over all months. The monthly utilities can be seen as measures of the patients quality adjusted life month (QAL month). Based on the judgment of treating physicians and the patients pain assessment reports, the utilities were assumed for each state (Table 1.4). Without loss of generality, the reward for the respond state was rescaled to be 1. Receiving a reward of 1 indicates that the patient lived one month in the health state of responding to treatment for advanced cervical cancer. When the patient moves to a worse health state, the life quality is adjusted

Table 1.3: Transition probabilities of going to new health state given that a patient was in a prior health state at the end of the previous month (i.e., cycle).

FROM i to j		Respond	Limited Complications	Progress	Severe Complications	Die
Chemo only	Respond	0.8671	0.0024	0.127	0.0035	0
	Limited complications	1*	0*	0	0	0
	Progress	0	0	0.8623	0	0.1377
	Severe complications	0	0	0.9*	0.1*	0
	Die	0	0	0	0	1
Chemo + Bevacizumab	Respond	0.872	0.0273	0.0823	0.0184	0
	Limited complications	1*	0*	0	0	0
	Progress	0	0	0.8771	0	0.1229
	Severe complications	0	0	0.9*	0.1*	0
	Die	0	0	0	0	1

Note that the probabilities in a row must sum to 1, since 100% of the patients will either: 1) move from that health state to a new health state, or 2) remain in the same health state.
*probability is assumed by authors based on judgment of treating physicians.

downward for that month. The health utilities in Table 3 are similar to those used in the

Table 1.4: Health utilities assignments

HEALTH STATES	Respond	Progress	Limited Complications	Severe Complications	Die
UTILITY per month in state	1	0.5	0.75	0.5	0

Markov analysis by Refaat et al [10] to examine the use of bevacizumab for breast cancer treatment. One important difference is that Refaat et al assigned 0.25 for complications and we divided complications into severe (0.5 utility) and limited (0.75 utility).

1.3. Results

1.3.1 Estimating Cost

Based on the cost of treatment and medications to treat complications, the data involving cost/month were generated (Table 1.1). Once again this assumes, due to anticipated treatment delays, that each cycle is set to last for 1 month.

1.3.2 Markov Modeling

The cost-effectiveness model was developed using response, progression, and survival data from GOG 240 and the incidence of bevacizumab-specific complications as reported in the primary publication [3] along with the updated data [5,6]. Assignment of health utilities and probability estimation of time spent in one or another health state led to the construction of the Markov Decision Tree (Panel A in appendix).

1.3.3 Measuring Internal Validity of the Markov Model

As described in the preceding section, to be able to describe the gains in survival time in expected life months, the Markov model was simplified by having each treatment cycle (and health status state) occur at 28-day intervals. We checked the validity of our model by comparing with the primary manuscript, which reported an OS difference of 3.7 months favoring the arms that administered chemotherapy plus bevacizumab (17 versus 13.3 months), as well as the updated median of 3.9 months in the FDA approval [5]. In our Markov model, the expected life months until death were calculated to be 15 months for chemotherapy alone and 18.5 months for chemotherapy plus bevacizumab, a difference of a mean of 3.5 months. Similarly, the difference in PFS also favors the patients receiving bevacizumab with 7.7 months for the chemotherapy alone cohort and 10.4 months for those who received chemotherapy plus bevacizumab, a difference of 2.7 months. In both analyses, and consistent with the findings of the original paper, treatment with chemotherapy plus bevacizumab yields higher expected life months.

1.3.4 Expected Cost and Cost Effectiveness

The estimated total cost of therapy with bevacizumab is approximately 13.2 times that for chemotherapy alone. Specifically, for each patient, the estimated total cost of chemotherapy alone is $6,053 and that of chemotherapy plus bevacizumab is $79,844. In terms of the OS advantage described by the Markov model, an average gain of 3.5 life months will cost an extra $73,791. Figure 1.2 depicts a tradeoff between life months gained and increased cost of therapy incorporating bevacizumab. The ICER is $21,083/month ($252,996/year). If the payer is able or willing to pay $21,083 for one more additional life month ($252,996 for one more additional life year) before death, then chemotherapy plus bevacizumab should be administered.

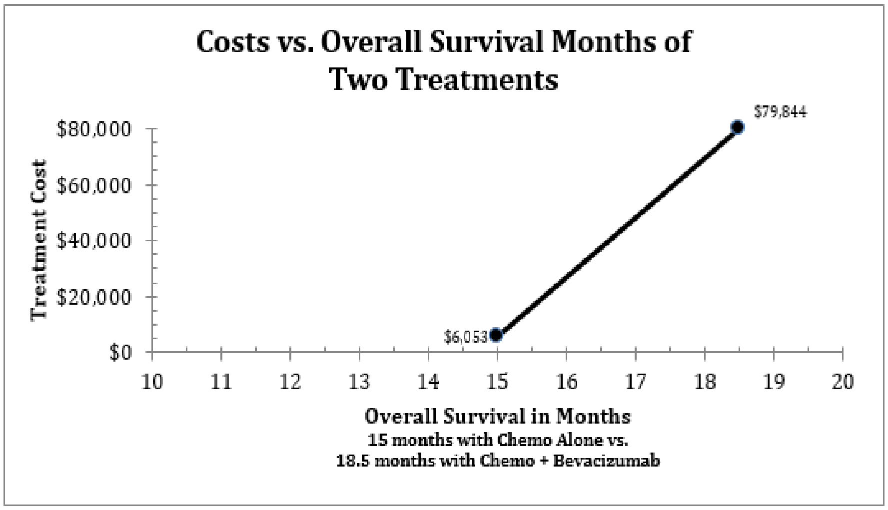


Figure 1.2: Cost effectiveness analysis of chemotherapy with and without bevacizumab in life months until death

Because treatment with chemotherapy plus bevacizumab leads to an increase in bevacizumab-specific complications, to better analyze cost-effectiveness, the impact of decrease in QoL from complications was modeled by QALmonths. For example, as specified in the model, the severe complication state yields a utility of 0.5 per cycle, compared with a 1 for a person in the respond state. The expected QALmonth for the chemotherapy plus bevacizumab cohort is higher than that for the chemotherapy alone cohort. When the QALmonth measure is used, the difference goes down to 3.0QALmonth (14.3-11.3 QALmonth or 0.25 QALY). The ICER increases to $24,597/QALmonth ($73,731/3 months or $73,731/(3/12) = $295,164/QALY) due to the smaller difference in QALmonths (see dashed line in Figure 1.3). For these patients, an increase of an average of 3.5 months alive (living in the different possible states (respond, progress, limited complications, or severe complications) is modeled as equivalent to 3.0 months in the respond state. A sensitivity analysis of remaining in the severe complication state for an additional month appears in Panel B in appendix.

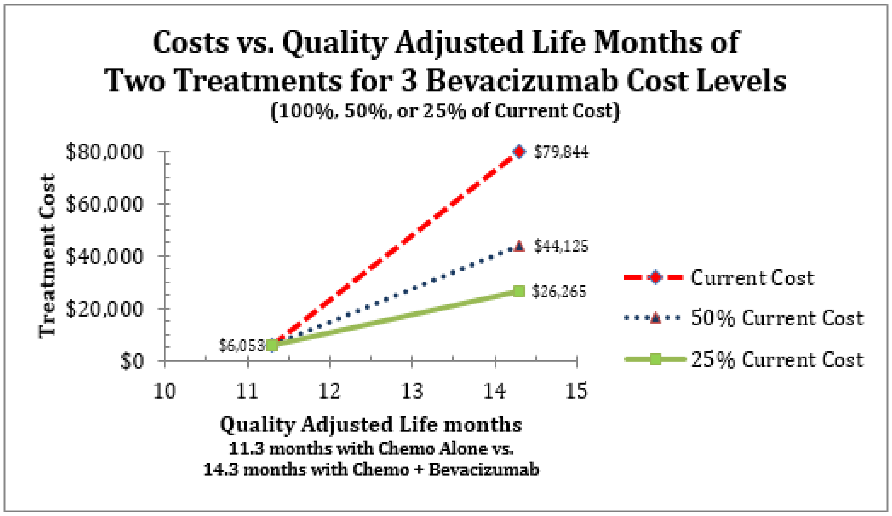


Figure 1.3: Cost-effectiveness analysis of chemotherapy with and without bevacizumab for QALmonths and with projected reduction in cost of bevacizumab.

1.3.5 Projected Impact of Decreasing the Cost of Bevacizumab

If the cost of bevacizumab were to decrease substantially, both the total cost of the chemotherapy plus bevacizumab treatment and the ICER will be reduced without change in efficacy (Figure 1.3). With a 50% reduction in the cost of bevacizumab, the ICER is $12,691/QALmonth ($152,292/QALY). This translates to $38,072 for the 3.5 month (or 0.29 year) gain in OS. With a reduction to only 25% of current cost, the ICER is $6737/QALmonth ($80,844/QALY). This translates to $23,580 for the 3.5 month (or 0.29 year) gain in OS.

1.4. Discussion

One of the major challenges facing healthcare worldwide is the incremental cost-effectiveness and the threshold for using or rejecting specific drugs. Bevacizumab is one of the most expensive drugs currently available. In many countries, its use has been restricted based on cost-effectiveness studies that suggest that the drug is not cost-effective.

Cost-effectiveness studies on integrating bevacizumab in the management of US FDA-approved indications have been performed previously and include metastatic and recurrent colorectal cancer, primary untreated non-small cell lung cancer, and renal cell carcinoma [13-20]. For example, Shiroiwa et al reported a maximum ICER of $145,000 per life year for colorectal cancer [16], while Chien et al reported a maximum ICER of over $300,000 for patients with non-small cell lung cancer for whom bevacizumab was added to chemotherapy [18]. Although in their economic evaluation of new targeted therapies Benedict et al did not report the ICER for bevacizumab in the treatment of metastatic renal cell carcinoma, the investigators concluded that sunitinib is a cost-effective alternative to befacizuamb with savings of $67,798 per patient treated in the United States [20].

Although approved for recurrent glioblastoma, cost-effectiveness studies for this indication are lacking [22-25]. Additionally, cost-effectiveness of bevacizumab in metastatic breast cancer has been evaluated [10-12, 21], with marginal cost effectiveness of $232,720.72 reported by Refaat et al. [10]. Although not approved in age-related macular degeneration, bevacizumab is considered an acceptable alternative to ranibizumab based on a randomized trial [26].

In four phase III randomized studies in newly diagnosed, platinum sensitive, and platinum resistant ovarian cancer, the arms administering chemotherapy and bevacizumab all met their primary endpoints with significant improvements in PFS [27-30]. Bevacizumab has not been approved in the U.S. for frontline ovarian cancer therapy, although the FDA has approved use in patients with platinum-resistant recurrent disease. Cohn et al. evaluated GOG 218, which studied bevacizumab in frontline therapy and concluded that the addition of bevacizumab to standard chemotherapy was not cost-effective with an ICER of $401,088 per progression-free life year saved for the bevacizumab throughout arm (primary plus maintenance therapy) [31]. The ICER fell below $100,000 per progression-free life year saved when the cost of bevacizumab was reduced to 25 % of baseline. In another cost-effectiveness analysis, Chan et al. reported that for the high risk subset from the ICON 7 study that experienced an OS benefit, the incremental cost of bevacizumab was $170,000 [31-32].

The dominant theme to emerge from cost-effectiveness studies is that with the exception of the non-lethal condition of age-related macular degeneration for which very small dosages of drug are required, bevacizumab will not be cost-effective in the management of solid tumor malignancies due to the current high cost of the drug, relatively limited impact on duration of survival, and healthcare expenditures required to manage anti-VEGF-specific toxicology [10-12, 15, 16, 18, 20, 31-33].

The growing use of bevacizumab can be demonstrated by sales data. In 2013, with global sales of $6.7 billion, bevacizumab ranked 9th in terms of revenue generated among the top 50 pharmaceutical agents [34]. Looking back, sales for bevacizumab grew by 9% between 2011 and 2012 to reach $6.3 billion in 2012 compared to $5.8 billion in 2011 [35]. The increase was attributable to increased usage in established indications (colorectal and lung cancer), along with E.U. approval to treat platinum-sensitive ovarian cancer which was granted in 2012. In the U.S. market sales of bevacizumab increased from $2.6 billion in 2011 to $2.7 billion in 2012 while in Western Europe sales increased from $1.6 billion in 2011 and $1.7 billion in 2012 [35]. Sales in other international markets were boosted by the CEMAI region (Central and Eastern Europe, Middle East, Africa, and the Indian Subcontinent), Latin America, and the APAC regions (Australia, China, and Japan). Based on growing sales, health care payers have implicitly indicated a relatively high willingness to pay per QALY gained.

For our analysis we chose a monthly Markov cycle because that time period corresponds with the time span in which a patient could transition to a new health state. Furthermore, because survival for patients with advanced cervical cancer is measured in months rather than years, we feel that our choice of reporting results in QALmonth is appropriate, adjusted from a baseline of living a month responding to treatment for advanced cervical cancer. We found that the cost of therapy resulting from the incorporation of bevacizumab was nearly 13.2 times that of chemotherapy alone and when taking into account complications, the ICER is $24,597/QALmonth (or $295,164/QALY) or a mean of $73,791 extra for a single patient over the course of the treatment, over the cost of chemotherapy alone.

Investigators bringing bevacizumab to cervical cancer are not in a position to determine whether $73,791 cost per patient treated is cost-effective therapy. Similarly, those studying cost of care are unable to assign a price to a gain in 3.9 months of a womans life. This is for society to determine. But what must be emphasized is that, with the exception of imatinib in chronic myelogenous leukemia, significant breakthroughs in oncology are currently not expected to impact survival beyond several months [36]. As a result, the ICERs associated with novel therapies may appear unacceptably high. The benefit conferred by bevacizumab to women with advanced cervical cancer is noteworthy as these cancers do not appear to be as chemosensitive as other solid tumors (eg., ovarian cancer, etc). In addition, the population with recurrent/metastatic disease is unique as the majority have been previously irradiated which leads to diminished bone marrow reserves and an increased risk for fistula formation. The FDAs August 14, 2014 decision to approve bevacizumab for advanced cervical cancer [4] constitutes a regulatory milestone allowing the study of potentially more efficacious treatments for cervical cancer to move forward.

The current study was limited by creation of a cost-effectiveness model from a singular data set, as GOG study 240 is the only randomized controlled clinical trial evaluating bevacizumab with chemotherapy in advanced cervical cancer [3]. Subsequently, this model does not incorporate all possible clinical outcomes. However, with recent FDA approval of this agent in advanced cervical cancer, the authors hope to repeat an analysis based on real world experience. Additionally, regarding potential costs of bevacizumab-related complications such as hypertension, fistula, thromboembolism or hemorrhage, there is limited information for which these costs were derived. These limitations may be addressed with further studies. Our model does not incorporate the societal impact of lost of productivity. Nor is this study from a patient perspective therefore, cost beyond therapeutic cost are not included. Finally, reimbursements and costs differ according to country and time making this analysis most relevant to 2013 and the United States.

While awaiting reform of the U.S. healthcare system and cost reconciliation, it appears that many cancer patients in need of oncologically effective but cost-ineffective therapies will be treated using the old arsenal of cytotoxic agents, an armamentarium of oncologic dead ends. When considering the relatively young median age at diagnosis of women with advanced cervical cancer, the number of life-years lost to family and to society are unacceptable. The societal and clinical dilemma can be reconciled from the vantage point of seeing things in the long-term. Specifically, with significant reductions in drug cost, the ICERs become more acceptable. This may be realized through the introduction of generics into the market.

Until reparations can be made to the broken U.S. healthcare system, it appears that many cancer patients in need of oncologically effective but cost-ineffective therapies will be treated using the old arsenal of cytotoxic agents, an armamentarium of oncologic dead ends. When considering the relatively young median age at diagnosis of women with advanced cervical cancer, the number of life-years lost to family and to society are unacceptable. The societal and clinical dilemma can be reconciled from the vantage point of seeing things in the long-term. Specifically, with significant reductions in drug cost, the ICERs become more acceptable. This may be realized through the introduction of generics into the market.

Biosimilars have been available on the European market since 2006 [37]. On June 27, 2013, the EMAs Committee for Medicinal Products for Human Use recommended approval for two biosimilar infliximab products to be marketed in the EU, making them the first biosimilar antibodies made available to patients in a highly regulated market [38].

As discussed above, our novel method of reporting in terms of QALmonths was intentional and indicative of how survival in advanced cervical cancer is measured. Our results can readily be translated to years to allow comparisons with other studies. For example, when there is a reduced cost of bevacizumab to 25% of the current baseline cost, the ICER is $80,844/QALyear.This suggests that perhaps through the availability of biosimilars, anti-VEGF therapy can be cost-effective in advanced cervical cancer.

1.5. Bibliography

[1] Monk BJ, Sill MW, McMeekin DS, Cohn DE, et al: Phase III trial of four cisplatin-containing doublet combinations in stage IVB, recurrent, or persistent cervical carcinoma: A Gynecologic Oncology Group study. J Clin Oncol 27:4649-55, 2009.

[2] Monk BJ, Sill MW, Burger RA, et al: Phase II trial of bevacizumab in the treatment of persistent or recurrent squamous cell carcinoma of the cervix: A Gynecologic Oncology Group study. J Clin Oncol 27:1069-74, 2009.

[3] Tewari KS, Sill MW, Long HJ 3rd, et al: Improved survival with bevacizumab in advanced cervical cancer. N Engl J Med 370:734-43, 2014.

[4] Penson RT, Huang HQ, Tewari KS, et al: Patient reported outcomes in a randomized trial of bevacizumab in the treatment of advanced cervical cancer: A Gynecologic Oncology Group study. Lancet Oncol 2015 (in press).

[5] FDA News Release. FDA approves Avastin to treat patients with aggressive and late-stage cervical cancer. August 14, 2014. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm410121.htm.

[6] Tewari KS, Sill MW, Penson RT, et al. Final protocol-specified overall survival analysis of the phase III randomized trial of chemotherapy with and without bevacizuamb for advanced cervical cancer. Annual Meeting, European Society of Medical Oncology, Madrid, Spain, Late-Breaking Abstract 26: May 29, 2014.

[7] Phippen NT, Leath III CA, Havrilesky LJ, Barnett JC. Bevacizumab in recurrent, persistent or advanced stage carcinoma of the cervix: Is it cost-effective? Gynecol Oncol 2015;136:43-7.

[8] Shachtman RH, Schoenfelder JR, Hogue CJ: Conditional rate derivation in the presence of intervening variables using a Markov chain. Oper Res 30:1070-81, 1982.

[9] Standfleid L, Comans T, Scuffham P: Markov modeling and discrete event simulation in health care: A systematic comparison. Int J Technl Assess Health Care 30:165-72, 2014.

[10] Refaat T, Choi M, Gaber G, et al: Markov model and cost-effectiveness analysis of bevacizumab in HER2-negative metastatic breast cancer. Am J Clin Oncol 37:480-85, 2013.

[11] Montero AJ, Avancha K, Glck S et al: A cost-benefit analysis of bevacizumab in combination with paclitaxel in the first-line treatment of patients with metastatic breast cancer. Breast Cancer Research and Treatment 132: 747-51, 2012.

[12] Lloyd A, Nafees B, Narewska J, et al: Health state utilities for metastatic breast cancer. British Journal of Cancer 95: 683-90, 2006.

[13] Hurwitz H, Fehrenbacher L, Novotny W, et al: Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med 350:2335-42, 2004.

[14] Giantonio BJ, Catalano PJ, Meropol NJ, et al: Eastern Cooperative Oncology Group Study E3200. Bevacizumab in combination with oxaliplatin, fluorouracil, and leucovorin (FOLFOX4) for previously treated metastatic colorectal cancer: Results from the Eastern Cooperative Oncology Group Study E3200. J Clin Oncol 25:1539-44, 2007.

[15] Ruiz-Millo O, Albert-Mari A, Sendra-Garcia A, et al: Comparative cost-effectiveness of bevacizumab-irinotecan-fluorouracil versus irinotecan-fluorouracil in first-line metastatic colorectal cancer. J Oncol Pharm Pract 20:341-50, 2014.

[16] Shiroiwa T, Fukuda T, Tsutani K: Cost-effectiveness analysis of bevaczziumab combined with chemotherapy for the treatment of metastatic colorectal cancer in Japan. Clin Ther 29:2256-67, 2007.

[17] Sandler A, Gray R, Perry MC: Paclitaxel-carboplatin alone or with bevacizumab for non-small cell lung cancer. N Engl J Med 355:2542-40, 2006.

[18] Chien CR, Shih YC: Economic evaluation of bevacizumab in the treatment of non-small cell lung cancer (NSCLC). Clinicoecon Outcomes Res 4:201-8, 2012.

[19] Escudier B, Pluzanska A, Koralewski P, et al: AVOREN Trial investigators. Bevacizumab plus interferon alfa-2a for treatment of metastatic renal cell carcinoma: A randomized, double-blind phase III trial. Lancet 370:2103-11, 2007.

[20] Benedict A, Figlin RA, Sandstrom P: Economic evaluation of new targeted therapies for the first-line treatment of patients with metastatic renal cell carcinoma. BJU Int 108:665-72, 2011.

[21] Miller K, Wang M, Gralow J: Paclitaxel plus bevaczumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med 357:2666-76, 2007.

[22] Friedman HS, Prados MD, Wen PY, et al: Bevacizumab alone and in combination with irinotecan in recurrent glioblastoma. J Clin Oncol 27:4733-40, 2009.

[23] Kreisl TN, Kim L, Moore K, et al: Phase II trial of single agent bevacizumab followed by bevacizumab plus irinotecan at tumor progression in recurrent glioblastoma. J Clin Oncol 27:740-5, 2009.

[24] Gilbert MR, Dignam JJ, Armstrong TS, et al: A randomized trial of bevacizumab for newly diagnosed glioblastoma. N Engl J Med 370:699-708, 2014.

[25] Chinot OL, Wick W, Mason W, et al: Bevacziuamb plus radiotherapy-temozolomide for newly diagnosed glioblastoma. N Engl J Med 370:709-22, 2014.

[26] CATT Research Group, Martin DF, Maguire MG, Ying GS, et al. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. N Engl J Med 364:1897-908, 2011.

[27] Burger RA, Brady MF, Bookman MA, et al: Incorporation of bevacizumab in the primary treatment of ovarian cancer. Gynecologic Oncology Group. N Engl J Med 365:2473-83, 2011.

[28] Perren TJ, Swart AM, Pfisterer J, et al: ICON7 Investigators. A phase 3 trial of bevacizumab in ovarian cancer. N Engl J Med 365:2484-96, 2011.

[29] Aghajanian C, Blank SV, Goff BA, et al: OCEANS: A randomized, double-blind, placebo-controlled phase III trial of chemotherapy with or without bevacizumab in patients with platinum-sensitive recurrent epithelial ovarian, primary peritoneal, or fallopian tube cancer. J Clin Oncol 30:2039-45, 2012.

[30] Pujade-Lauraine E, Hilpert F, Weber B, et al: Bevacizumab combined with chemotherapy for platinum-resistant recurrent ovarian cancer: The AURELIA open-label randomized phase III trial. J Clin Oncol 32:1302-8, 2014.

[31] Cohn DE, Kim KH, Resnick KE, et al: At what cost does a potential survival advantage of bevacizumab make sense for the primary treatment of ovarian cancer? A cost-effectiveness analysis. J Clin Oncol 29:1247-51, 2011.

[32] Chan JK, Herzog TJ, Hu L, et al: Bevacizumab in treatment of high-risk ovarian cancer A cost-effectiveness analysis. Oncologist 19:523-7, 2014.

[33] Patel JJ, Mendes MA, Bounthavong M, et al: Cost-utility analysis of bevacizumab versus ranibizumab in neovascular age-related macular degeneration using a Markov model. J Eval Clin Pract 18:247-55, 2012.

[34] Goodman M. Market watch: Pharma industry strategic performance: 2007-2012E. Nat Rev Drug Discov 7:967, 2008.

[35] Roche Finance Report 2012. Jan 28, 2013 Roche Group Financial Review. Available at www.roche.com/fb123.pdf. Accessed on February 22, 2015.

[36] Kantarjian H, Sawyers C, Hochhaus A, et al. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. N Engl J Med 346:645-52, 2002.

[37] Ebbers HC, van Meer PJ, Moors EH et al. Measures of biosimilarity in monoclonal antibodies in oncology: The case of bevacizumab. Drug Discov Today 18:872-9, 2013.

[38] Beck Reichert JM. Approval of the first biosimilar antibodies in Europe: A major landmark for the biopharmaceutical industry. MAbs 5:621-3, 2013.

[39] Avastin (Bevacizumab by Roche). https://gmrdata.com/downloadable/download/linkSample/link_id/12

[40] Generics and Biosimilars Initiative, update. http://gabioline.net/Biosimilars/General/Biosimilars-of-bevacizumab

1.6. Appendix

The overall structure of the Markov tree demonstrates primary branching at the point of randomization between chemotherapy alone and chemotherapy plus bevacizumab. Note that the termination condition was set to 60 cycles which corresponds to 5 years. At this point in time, 99% of patients are expected to have died. (# sign indicates a probability = 1 the sum of the other probabilities following a chance node circle).

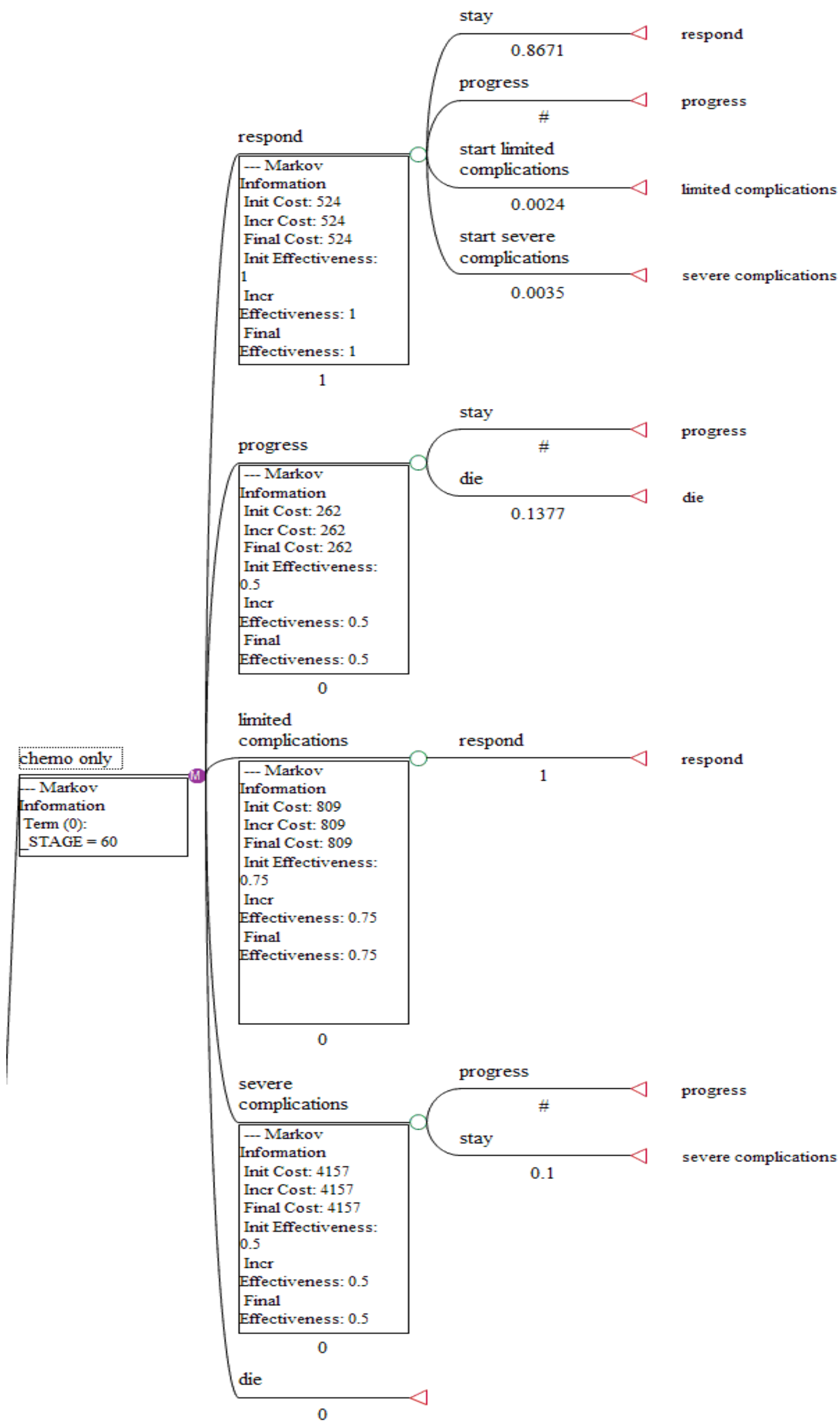


Figure 1.4: Panel A. Terminal branching of the chemotherapy alone cohort

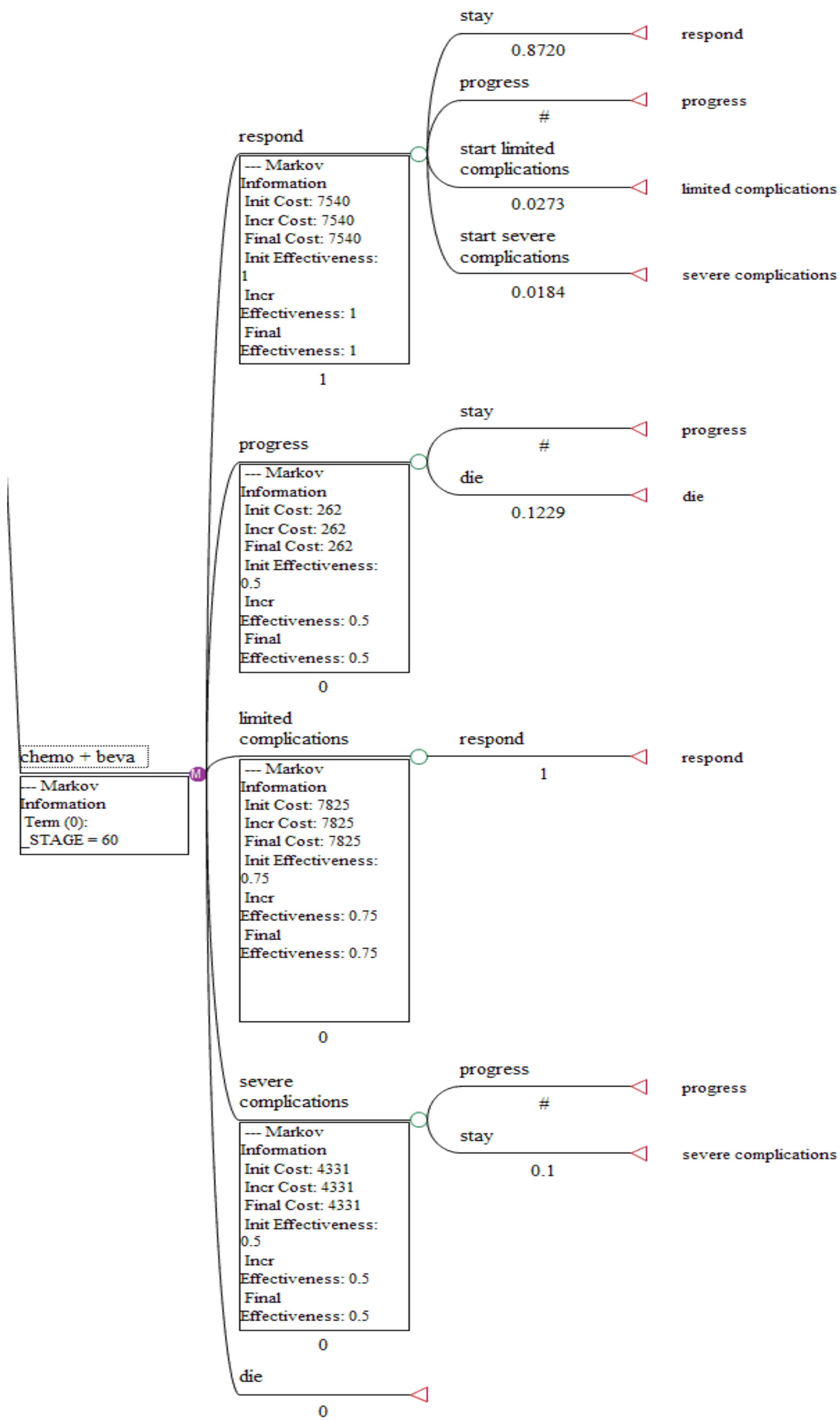
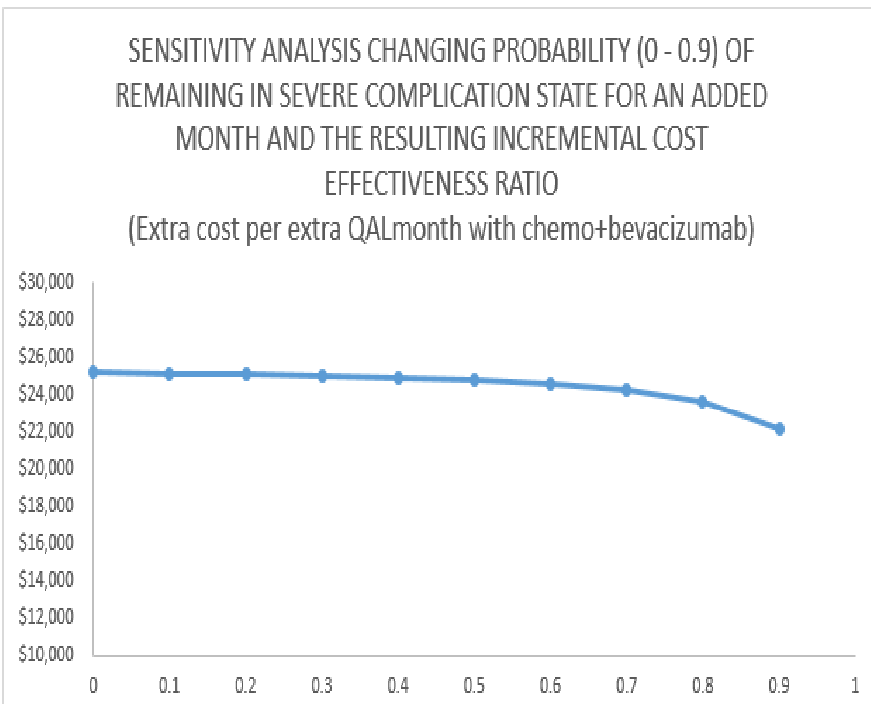


Figure 1.5: Panel B. Terminal branching of the chemotherapy plus bevacizumab cohort

We ran a sensitivity analysis on the transition probability of staying in the severe complication state an additional month. In the base case analysis, we assumed the probability was 0.1 for both treatments. The incremental cost effectiveness ratio (ICER), which is the added cost per added QALmonth of survival with chemotherapy + bevacizumab compared with chemotherapy alone, ranged from $25,176 to $22,155 as the transition probability ranged from 0 to 0.9. If a patient stays in the severe complication state for only one month, so the transition probability is 0, the ICER is $25,176 (it costs $25,176 extra per month of added survival with chemotherapy + bevacizumab). If the transition probability is 0.9, so it is very likely the patient stays in the severe complications state another month, the ICER is $22,155.



Chapter 2

Retail Distribution Strategy with Outlet Stores

2.1. Introduction

Outlet stores are brick-and-mortar stores that are usually located far away from city centers that provide older, less desirable, or lower quality products at deep discounted prices. Today, there are hundreds of outlet malls nationwide. Managing outlet stores has become an important channel for gaining high profits as they contribute to a large share of the total revenue. For example, the total revenue generated from outlet stores is \$45.6 billion in North America in 2015 (Humphers (2015)). Figure 1 indicates that the percentage of stores that are outlets is increasing over the years for a number of major retailing companies (Kapner (2014)). Recently, firms such as Macy's have closed some primary stores and opened more outlet stores instead, some of which are close to or at the same location of their primary stores. The objective is to lower costs and cater to customer demand for lower priced products (see, e.g. Kieler (2016), Zhang (2016)).

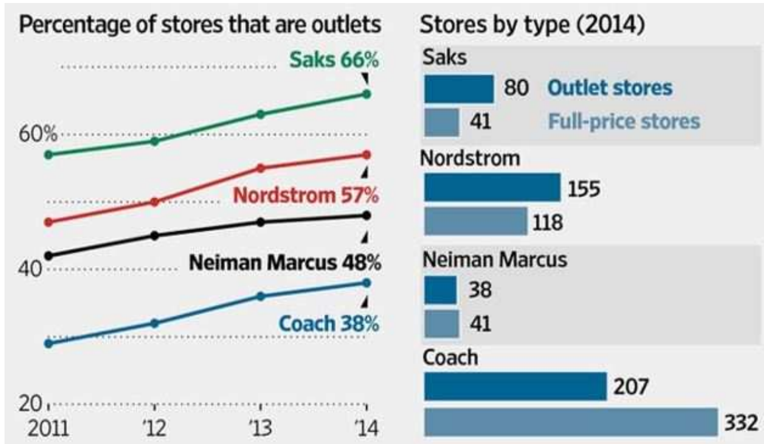


Figure 2.1: Percentage of Outlet Stores

There has been several reasons behind the existence of outlet stores. Traditionally, the main purpose of opening an outlet store is to dispose excess or out-of-date or damaged items that cannot be sold in the regular store. In view of the increasing traffic to the outlet stores, companies now even have planned overstocks to sell in the outlet stores, usually with discount prices, to attract lower-value customers with higher willingness to travel. For example, J.Crew, Gap and Saks Off 5th operate specific product lines that are dedicated to products sold in outlet stores (Maheshwari (2014)). There is also academic literature discussing the pros and cons of operating outlet stores (Coughlan and Soberman (2004), Ngwe (2014)). First, one of the main benefits is to price discriminate consumers. For instance, T.J. Maxx and Marshall use discounted prices to attract lower-valuation consumers (Rocha (2010)). Another reason is quality differentiation (Lieber (2014)). For example, Coach sells products with coach factory logo only in its outlet store (Brandculture (2010)). On the cost side, a direct benefit is to avoid high fixed costs as outlet stores are usually located far away from the city centers.

There are some concerns as well associated with operating outlet stores. Lower quality products sold in outlet stores may dilute the brand, resulting in a low-end brand image to the customers, even though the firm could actually improve its quality offering in the main store. Another concern is that outlet store may cannibalize sales in the main store, resulting in a reduction in its profit generated from the main store. However, location differentiation may alleviate the cannibalization effect. The prevalence of outlet stores and the vast qualitative discussions and empirical studies on their existence motivate us to examine this interesting problem from a theoretical perspective. We develop an analytical model aiming to incorporate the most common benefits and concerns related to outlet stores. More specifically, we would like to address the following main research questions based on our model:

(1) Under what conditions would the firm operate an outlet store, in addition to the main store?

(2) If an outlet store is offered, how would the firm differentiate the two stores in terms of product quality, prices, and store location?

(3) Finally, how does the offering of an outlet store affect the main store in terms of its sales, product price and quality?

To answer these questions, we consider a model with one firm who considers offering an outlet store in addition to its existing main store. We consider two types of market structures. In the first structure, the firm only operates the main store, whereas in the second structure, both the main and outlet stores are operated. The firm distributes regular products in the main store, but sells lower-quality products through the outlet channel. Both stores compete for the same set of consumers who are located near the main store but are vertically differentiated in terms of their valuation of the product. If an outlet store is available and consumers decide to buy from there, they need to travel to the outlet store with a certain disutility associated with travel. There are three kinds of differentiation which jointly drive the equilibrium characterization – location differentiation based on the disutility of traveling to the outlet store, product differentiation measured by the difference in product quality, and price differentiation between products sold at the main and outlet stores. The sequence of events is as follows: In the first stage of the game, the firm chooses whether to operate the outlet store. In the second stage, the firm decides the corresponding product quality and price at the main store, and if the outlet store is used, the firm also sets price, quality and location for the outlet store. In the third stage, consumers decide whether and where to buy.

Our base model is evaluated under a setting where consumers are heterogeneous in their valuation of the product and are equally sensitive to travel to the outlet store (irrespective of their valuation of the product). We characterize the equilibrium strategy on whether or not to offer an outlet store and the corresponding operational decisions. The key findings are as follows: i) The presence of the outlet store would actually encourage the firm to improve the product quality in the main store. This is consistent with the empirical finding in Ngwe (2014). ii) When the fixed cost or consumers' travel sensitivity or base unit production cost is lower, it is more profitable for the firm to operate an outlet store. iii) Product quality, price and location differentiation are not necessarily monotone in travel sensitivity. iv) Finally, higher location differentiation between the main and outlet stores leads to lower differentiation in quality and prices. We also examine the effect of e-commerce on the firm's optimal strategy. As competition from e-commerce increases, we show that the relevance of outlet stores becomes stronger – not only are they located closer to the main store, their share of total sales is higher. Conversely, sales at the main stores declines which is consistent with the anecdotal evidence.

Finally, we extend our analysis to a setting where consumers are heterogeneous in their travel sensitivity: low-valuation customers are more willing to travel to the outlet store, relative to the high-valuation customers. Our analysis indicates that the more different customers are in terms of their willingness to travel, the more likely is it for the firm to adopt a segmentation strategy in the sense that the main store will be operated to target high-valuation customers only and the outlet store will be operated to target low-valuation customers only. Another result is that an increase in the high-valuation customers' travel sensitivity may increase the firm's profit. This is in contrast with the previous case where profits are lower with high travel sensitivity.

The rest of the paper is organized as follows. We review the related literature in §2. In §3, we introduce the base model. In §4, we characterize the optimal store offering strategy, price, quality and location decisions and also discuss their implications. In the end of this section we explore the effect of e-commerce. In §5, we extend our model to the situation when the consumers are heterogeneous in travel sensitivity, and finally we conclude in §6.

2.2. Literature Review

There is limited literature studying operational problems in outlet stores. The most relevant literature to our paper is on vertical line extensions in multi-channel retailing, especially when one of the channels is in the form of an outlet/factory store which, in general, is geographically distanced from the city center (Coughlan and Soberman (2005)). There are many justifications behind vertical line extension strategies in the literature. For example, the introduction of a new channel, which could be differentiated and geographically distanced from the existing channel(s), may reach more customer bases (Kekre et al. (1990)), provide advertising effects (see, e.g., Qian et al. (2013)), and serve as a competing tool for the supplier with its downstream and independent retailers (see, e.g., Bell et al. (2003) and Chiang et al. (2003)). More often, such a new channel can be used to better segment the existing customer groups (see, e.g., Coughlan and Soberman (2005), Ngwe (2014), and Cao et al. (2016)). Our paper falls in this category. In particular, we consider *whether or not* the firm should introduce an outlet store selling lower-quality products and *where* the store should be located. See, e.g., Manez et al. (2001) for a detailed review on papers about vertical line extensions.

There are two recent empirical papers that focus on the outlet stores: Qian et al. (2013) and Ngwe (2014). In the former, the focus is on the positive spillovers of the outlet store on the main channels due to the advertising effect caused by the outlet store. We consider a single period model here so no advertising effect is incorporated. The empirical study by Ngwe (2014) shows a positive correlation between consumers' valuation of the product and their sensitivity to travel distance. Our extended model with heterogeneous travel sensitivity provides an analytical prediction on how this positive correlation affects the firm's decisions. In addition, our analytical result supports Ngwe (2014)'s empirical observation that the introduction of an outlet store would improve the product quality at the primary store.

There are an extensive list of analytical papers focusing on the optimal product line pricing and design (Villas-Boas (1998)). See, e.g., Tsay et al. (2004) for a detailed review of this literature. The main issue in this stream is about how to price discriminate the products sold via different channels to mitigate the cannibalization effect of these channels. See, e.g., Mussa and Rosen (1978), Moorthy (1988), Deneckere and McAfee (1996), Bernstein et al. (2004), and Anderson and Dana (2009). The firm in our paper also faces the cannibalization issue when considering a new outlet store. It can reduce this effect via not only pricing, but also quality *and* location differentiation.

Finally, all the papers discussed above assume that products/lines are vertically differentiated. There are papers that study horizontally differentiated products. The relevant issues in this case could be decisions involved in product assortment (see, e.g., Ryzin et al. (1999) and Kok et al. (2007)). Our paper considers vertically differentiated products and lines.

In addition, we extend our model to study the effect of e-commerce. We find the current new phenomenon that companies are closing main stores and opening outlet stores with a closer distance to city centers (see, e.g. Kieler (2016), Zhang (2016)) can be explained by the growing popularity of online shopping.

2.3. Model Framework

Consider a firm that currently operates a store in the city center where the customer base is located. We name it as the "main" store. The firm is now considering whether or not to open an "outlet" store which could differ from the main store in several possible dimensions – quality, price and location. As a result, customers may experience a different utility when buying from these stores. Specifically, these utilities can be expressed as:

$$U_m = \delta_m v - p_m \text{ and } U_o = \delta_o v - at - p_o,$$

where:

m and o: index for the main and outlet stores, respectively; p_m and p_o: selling prices in the main and outlet stores, respectively; v: base valuation that a customer can obtain from consuming the product; δ_m and δ_o: quality of products sold in the main and outlet stores, respectively; t: distance of the outlet store away from the main store in the city center.

To capture customers' heterogeneity in their valuation of the product, we assume that v follows a uniform distribution on $[0, 1]$. We also assume that the product sold in the main store has a higher quality level than that in the outlet store, that is, $\delta_m \geq \delta_o \geq 0$. Since the outlet store is t distance away from the city center where the customers are, there is a disutility, measured by at, if customers decide to travel to the outlet store. Given this formulation, we observe three types of differentiation between the two stores: (1) *price* differentiation measured by $p_m - p_o$; (2) *quality* differentiation quantified by $\delta_m - \delta_o$; and finally (3) *location* differentiation represented by t. Due to the support of v being on $[0, 1]$, we also normalize the range of a and t to be in $[0, 1]$.

Note that the firm may decide to locate the outlet store at the same site of the main store, in which case, we have $t = 0$. In this paper, all of the three types of differentiation between the main and outlet stores are formulated as decision variables. Given these decisions, consumers would choose whether or not to buy and where to buy based on the comparison of the net utilities from not buying, buying from the main store, and buying from the outlet store (if available).

From the firm's perspective, an outlet store may offer potentially lower quality products at lower prices. This could encourage customers who originally did not buy from the main store (due to their low valuations) to now buy from the outlet store. A broader market coverage could help the firm to gain more revenue. This benefit then needs to be compared with the costs associated with operating the outlet store. For the main store, we use c_m and F_m to denote the unit production cost and fixed cost, respectively. Since we assume that the firm makes a decision on whether or not to open an outlet store in addition to the main store, the fixed cost of the main store, F_m, is sunk and will be ignored in the subsequent profit function. For the outlet store, we use c_o to represent the unit production cost. However, its fixed cost depends on the location of the outlet store, denoted by $(1-t)^2F_o$, where F_o can be interpreted as the fixed cost of operating the outlet store at the main store site. The quadratic (convex) fixed cost in location implies a non linear effect on costs with travel distance. Since the product in the main store has higher quality compared to the outlet store, that is, $\delta_m \geq \delta_o$, it is natural to also assume that $c_m \geq c_o$. To simplify, we further assume that $c_m = \delta_m^2 c$ and $c_o = \delta_o^2 c$. The quadratic cost of quality implies that the higher the quality is, the more costly it is to improve the quality further. Similar quality-dependent cost functions are widely used in the economics, marketing and operations literature, see, e.g, Banker et al. (1998), Matsubayashi (2007). Again, due to the assumption that $v \sim U[0,1]$, we restrict F_m, F_o and c to be in $[0,1]$ as well. Taking into account the consumers' choice model, the firm will first decide whether or not to introduce a new outlet store. If an outlet store is introduced, the firm then sets the level of quality, price and location differentiation between the main and outlet stores. In what follows, we will characterize the demand and profit functions. For convenience, the notation is summarized in Table 1.

Table 2.1: Summary of Notation

Symbol	Definition
m	main store
o	outlet store
c	base production cost, where $c \in [0, 1]$
a	consumer sensitivity to travel cost
F_m	fixed cost for the main store, which is sunk
F_o	maximum fixed cost for the outlet store, where $F_o \in [0, 1]$
v	product valuation, where $v \sim U[0, 1]$
t	travel distance to the outlet store, decision variable, where $t \in [0, 1]$
δ_m	product quality in the main store, decision variable, where $\delta_m \in [0, \infty)$
δ_o	product quality in the outlet store, decision variable, where $\delta_o \in [0, \delta_m]$
p_m	price in the main store, decision variable, where $p_m \in [0, \delta_m]$
p_o	price in the outlet store, decision variable, where $p_o \in [0, \delta_o]$
Q_m	demand in the main store
Q_o	demand in the outlet store
Π^m	profit with only the main store
Π^{mo}	profit with both the main and outlet stores

We first consider the simple case without an outlet store. A customer would buy from the main store if $U_m = \delta_m v - p_m \geq 0$. Hence, customers with valuation in $[\frac{p_m}{\delta_m}, 1]$ would buy the product and the demand is $Q_m = 1 - \frac{p_m}{\delta_m}$, where $0 \leq p_m \leq \delta_m$. The firm's corresponding profit function is:

$$\Pi^m = (p_m - c_m)Q_m = (p_m - \delta_m^2 c)(1 - \frac{p_m}{\delta_m}), \tag{2.1}$$

where δ_m and p_m are the quality and pricing decisions.

In the case with both the main and outlet stores, we have location differentiation in addition to quality and price differentiation. A customer would buy from the main store if $U_m = \delta_m v - p_m \geq U_o = \delta_o v - at - p_o$ and $U_m \geq 0$. The first inequality yields an indifference point $v_3 = \frac{p_m - p_o - at}{\delta_m - \delta_o}$, and the second inequality yields another indifference point $v_1 = \frac{p_m}{\delta_m}$. Similarly, a customer would buy from the outlet store if $U_m = \delta_m v - p_m \leq U_o = \delta_o v - at - p_o$ and $U_o \geq 0$. The indifference point from the first inequality is again v_3 and the second inequality yields another indifference point $v_2 = \frac{p_o + at}{\delta_o}$. To derive demand for each store, we need to further compare the three indifference points, v_1, v_2 and v_3, which leads to three possible cases.

(1) If $v_1 \leq v_2$, we always have $v_3 \leq v_1 \leq v_2$. In this case, consumers with valuation in $[v_1, 1]$ would buy from the main store, the rest would buy nothing. This case is equivalent to the model without an outlet store. Hence, we call it as case (m).

(2) If $v_2 \leq v_1$ and $v_3 \leq 1$, we always have $v_2 \leq v_1 \leq v_3$. Hence, consumers in $[v_3, 1]$ would buy from the main store, and those in $[v_2, v_3]$ prefer the outlet store and the rest choose not to buy. Since there is positive sales in both stores, we call this case as case (mo).

(3) If $v_2 \leq v_1$ and $v_3 \geq 1$, consumers in $[v_2, 1]$ buy from the outlet store and the rest do not buy. Since there is positive sales only in the outlet store, we call this case as case (o).

The demand functions can be summarized as follows:

$$(Q_m, Q_o) = \begin{cases} (1 - \frac{p_m}{\delta_m}, & 0) & \text{if in case } (m) \\ (1 - \frac{p_m - p_o - at}{\delta_m - \delta_o}, & \frac{p_m - p_o - at}{\delta_m - \delta_o} - \frac{p_o + at}{\delta_o}) & \text{if in case } (mo) \\ (0, & 1 - \frac{p_o + at}{\delta_o}) & \text{if in case } (o), \end{cases}$$

where case (m) implies that sales occurs in only the main store when $p_o \geq \frac{p_m \delta_o}{\delta_m} - at$; case (mo) implies that sales occurs in both stores when $p_m - at + \delta_o - \delta_m < p_o < \frac{p_m \delta_o}{\delta_m} - at$; and case (o) implies that sales occurs in only the outlet store when $p_o \leq p_m - at + \delta_o - \delta_m$. Given the demand functions for the stores, we can express the firm's total profit function as follows:

$$\begin{aligned} \Pi^{mo} &= (p_m - c_m)Q_m + (p_o - c_o)Q_o - (1-t)^2 F_o, \qquad (2.2) \\ &= (p_m - \delta_m^2 c)Q_m + (p_o - \delta_o^2 c)Q_o - (1-t)^2 F_o, \end{aligned}$$

where $p_o, p_m, \delta_o, \delta_m$ and t are the decisions for the firm. Note that the firm's total profit is the sum of its profits generated from both the main and the outlet stores; i.e., $\Pi^{mo} = \Pi^m + \Pi^o$. It is possible that the sales and hence the profit of one store is zero in which case the total profit is simply the profit from the other store that has positive sales.

2.4. Model Analysis

Note from the demand functions that the model with an outlet store is more general than the model without. Hence, in this section, we will analyze the model with an outlet store. The analysis will provide insights into the conditions under which the firm would introduce the outlet store and how model parameters impact this decision. We first solve for the constrained optimal decisions in each of the three cases (i.e., cases (m), (mo) and (o)) in the demand function presented earlier, and then compare across these cases to identify the global optimal store strategy and the corresponding decisions. We start with the analysis of the constrained optimal decisions and profits in the three cases.

2.4.1 Equilibrium Analysis of Store Strategy

In order to derive the equilibrium, we first obtain the optimal decisions in each of the three constrained cases. The proofs of all the analytical results of §§4.1 and 4.2 are presented in Appendix A (Part I).

Lemma 1. *The constrained optimal decisions and profit in cases (m), (mo) and (o) are as follows:*

– Case (m): Positive sales exits only in the main store, where $(\delta_m = \frac{1}{3c},\ p_m = \frac{2}{9c})$*. Accordingly, the firm's profit is* $\Pi^{mo} = \frac{1}{27c}$*. For the outlet store,* (t, δ_o, p_o) *satisfies* $p_o \geq \frac{p_m \delta_o}{\delta_m} - at$*.*

– Case (mo): Positive sales exists in both the main and outlet stores, where:

- *If* $F_o \leq \min(\frac{a}{10}, \frac{2}{675c})$*:* $(\delta_m = \frac{2}{5c}, p_m = \frac{7}{25c})$*,* $(t = 0, \delta_o = \frac{1}{5c}, p_o = \frac{3}{25c})$ *and* $\Pi^{mo} = \frac{1}{25c} - F_o$*.*
- *If* $\frac{a}{10} \leq F_o \leq \bar{F}_o$*, where* $\bar{F}_o = 1$ *if* $a \leq \frac{1}{36c}$ *and* $\bar{F}_o = \frac{9(-3a^2c+\sqrt{15}\sqrt{-a^4c^2+36a^5c^3})}{8(-2+45ac)}$ *otherwise:* $(\delta_m = \frac{B}{5c}, p_m = \frac{-20A+7B}{50c})$*,* $(t = \frac{-9a^3c+2aF_o+10F_o^2-3\sqrt{9a^6c^2-4a^4cF_o+a^2F_o^2-20a^3cF_o^2}}{10F_o^2}, \delta_o = \frac{B}{10c}, p_o = \frac{3(-10A+B)}{50c})$ *and* $\Pi^{mo} = \frac{200A^2-10A+B}{25Bc} - (1-t)^2 F_o$*, where* $A = 20act$ *and* $B = 1 + \sqrt{1-20A}$*.*
- *Otherwise: It is not feasible to achieve positive sales in both stores.*

– Case (o): Positive sales exits only in the outlet store, where (δ_m, p_m) *satisfies* $p_o \leq p_m - at + \delta_o - \delta_m$*:*

- *If* $F_o \leq \frac{a}{6}$*:* $(t = 0, \delta_o = \frac{1}{3c}, p_o = \frac{2}{9c})$ *and* $\Pi^{mo} = \frac{1}{27c} - F_o$.

- *Otherwise:*
$(t = \frac{a^3c - 2aF_o + 6F_o^2 + \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}, \delta_o = \frac{1+\sqrt{1+12act}}{6c}, p_o = \frac{1-3act+\sqrt{1+12act}}{9c})$ *and* $\Pi^{mo} = \frac{(1-12act+\sqrt{1+12act})^2}{54c(1+\sqrt{1+12act})} - (1-t)^2F_o$.

The comparison of the firm's constrained optimal profits in the three cases presented in Lemma 1 leads to Proposition 1 below, which illustrates the firm's optimal store offering strategy. The corresponding global optimal decisions will be presented subsequently in Proposition 2.

Proposition 1. *The firm's optimal store offering strategy is characterized in Figure 2.2, where the cutoff curves are expressed as follows:*

L_a*:* $a = \frac{1}{36c}$ *if* $F_o \geq \frac{1}{288c}$*, and* $F_o = \frac{9(-3a^2c+\sqrt{15}\sqrt{-a^4c^2+36a^5c^3})}{8(-2+45ac)}$ *if* $\frac{2}{675c} \leq F_o \leq \frac{1}{288c}$*;*

L_b*:* $F_o = \frac{2}{675c}$ *if* $a \geq \frac{4}{135c}$*;*

L_c*:* $F_o = \frac{a}{10}$ *if* $a \leq \frac{4}{135c}$.

This result indicates that the firm would introduce an outlet store only if model parameters fall in either region R_2 or R_3, but not in region R_1 when the travel sensitivity a and the maximum fixed cost of the outlet store F_o is high. This seems intuitive since in region R_1, customers are not so willing to travel to the outlet store. In addition, offering such a store is also expensive as the fixed cost is high. When it is optimal to introduce the outlet store, the two stores would be operated in the same location in region R_2 and in separate locations in region R_3. Note that in region R_2, the fixed cost of operating such a store is low, even if it is kept at the same site of the main store. Since customers are not willing to travel a long distance to buy from the outlet store (as a is high), the firm would open the outlet store at the main store site. In region R_3, the fixed cost of operating the outlet store at the main

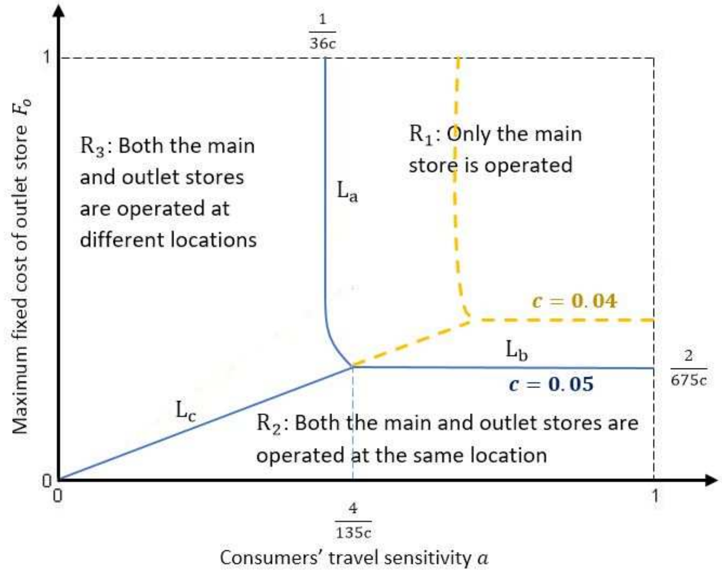


Figure 2.2: Firm's Optimal Store Offering Strategy

store site is high, the firm would move it away from the main store since travel sensitivity for customers is low. Finally, note that, when the unit base production cost, c, increases, size of region R_1 is larger, which implies that it is more likely for the firm to just operate the main store. This is because higher base unit production cost makes it more difficult for the firm to differentiate the products sold at the two stores.

The corresponding optimal decisions are given in the proposition below.

Proposition 2. *The firm's optimal decisions are summarized in Table* 2.2.

Table 2.2: Firm's Optimal Decision Variables

Region	R_1	R_2	R_3
Store offering strategy	Main store only	Both stores	Both stores
Outlet store location t	–	0	$\hat{t} \in [0,1]$
Main store quality δ_m	$\frac{1}{3c}$	$\frac{2}{5c}$	$\frac{B}{5c}$
Outlet store quality δ_o	–	$\frac{1}{5c}$	$\frac{B}{10c}$
Main store price p_m	$\frac{2}{9c}$	$\frac{7}{25c}$	$\frac{-20A+7B}{50c}$
Outlet store price p_o	–	$\frac{3}{25c}$	$\frac{3(-10A+B)}{50c}$
Main store demand Q_m	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{40A+B}{5B}$
Outlet store demand Q_o	–	$\frac{1}{5}$	$\frac{-60A+B}{5B}$
Main store profit	$\frac{1}{27c}$	$\frac{3}{125c}$	$\frac{(5-B)(40A+B)}{250c}$
Outlet store profit	–	$\frac{2}{125c} - F_o$	$\frac{(-60A+B)(-10A+B)}{125Bc} - (1-\hat{t})^2 F_o$

Note: $\hat{t} = \frac{-9a^3c+2aF_o+10F_o^2-3\sqrt{9a^6c^2-4a^4cF_o+a^2F_o^2-20a^3cF_o^2}}{10F_o^2}$, $A = ac\hat{t}$ and $B = 1+\sqrt{1-20A}$.

An interesting issue to study is how the presence of an outlet store would affect the firm's operational decisions in the main store. Note that if the firm decides, upfront, not to introduce the outlet store, the optimal decisions are the same as those in region R_1 in Table 2. Hence, the impact of the outlet store on the operational decisions can be obtained by comparing the optimal decisions in Table 2 and the values presented in region R_1 of the same table.

Proposition 3. *The existence of an outlet store would:*

(a) encourage the firm to improve the product quality in the main store, and

(b) lead to a higher selling price and hence lower sales and profit in the main store.

The rationale behind Proposition 3(a) is that cannibalization from the outlet store would force the firm to differentiate the two stores more on the quality (and price) perspective so as to maintain the competitiveness of the main store. This analytical result is supported by the empirical evidence presented in Ngwe (2014). Proposition 3(b) directly follows from Proposition 3(a) in that the profit generated at the main store is hurt by the outlet store due to channel cannibalization. That is, some customers may switch from the main store to the outlet store. Note that Qian et al. (2013) empirically shows a different result where demand and profit of the main store may increase due to the introduction of an outlet store. Their paper indicates that the positive impact is probably caused by the advertising effect of the outlet store. However, our paper does not take the advertising effect into consideration. Finally, even though the profit from the main store decreases due to the outlet store, the firm would overall benefit from it as the total profit is higher due to wider market coverage. This result is supported by both academic research and industry reports (see, e.g., Myers et al. (2004), Kumar and Venkatesan (2005) and Neslin and Shankar (2009)).

2.4.2 Sensitivity Analysis

Based on the optimal expressions presented in Proposition 2, we can also conduct some sensitivity analysis of the model parameters.

Corollary 1. *At optimality:*

(a) quality and price differentiation decreases in F_o and c; and

(b) location differentiation increases in F_o and c.

The implications of the results are as follows. First, when the fixed cost of operating the outlet store at the main store location is high, it is intuitive to create location differentiation in order to lower fixed costs. Due to the substitution effect of location differentiation, both quality and price differentiation will be low. Secondly, when the base unit production cost is high, it becomes difficult to adopt quality (and price) differentiation. So, the firm relies more on location differentiation.

Table 3 displays a complete summary of the sensitivity analysis of the store differentiation and profit with respect to model parameters, F_o, c and a.

Table 2.3: Sensitivity Analysis

	Fixed Cost F_o	Production Cost c	Distance Sensitivity a
Quality Diff.	$\downarrow$	$\downarrow$	not monotone
Price Diff.	$\downarrow$	$\downarrow$	not monotone
Location Diff.	$\uparrow$	$\uparrow$	not monotone
Profit	$\downarrow$	$\downarrow$	$\downarrow$

Following Table 3, we can draw a number of new implications in addition to Corollary 1. First, the effect of travel sensitivity a on store differentiation crucially depends on the value of fixed cost, F_o. For low values of F_o, location differentiation (t) is monotone and decreasing in travel sensitivity, which is as expected. However, when the fixed cost is high, location differentiation is no longer decreasing in a. This can be explained as follows. Recall that the firm uses three dimensions to differentiate the two stores: quality, price and location, and that the effective fixed cost of operating an outlet store is $(1-t)^2 F_o$. When F_o is low, as travel sensitivity a increases (or customers are less willing to travel), the firm can move the outlet store close enough to the main store (i.e., by reducing location t) without significantly increasing the effective fixed cost of the outlet store. However, when F_o is sufficiently high, using this strategy is quite costly, especially when the outlet store is close to the main store (i.e., when location differentiation (t) is small). So the firm would instead decrease the quality (and price) differentiation, which could lead to an increase in location t due to the substitution effect between these dimensions of differentiation observed in Corollary 1.

Second, the firm's profit would be negatively affected as the cost (either the fixed or unit production cost) associated with the outlet store increases or as customers become less willing to travel to the outlet store.

Lastly, note that from Table 2 that all the optimal decision variables can be expressed as a function of distance t. So, if we treat distance t as a parameter, in some sense, we can infer how the distance of the outlet store would affect the firm's price and quality decisions.

Corollary 2. *If we treat the location of the outlet store as fixed, that is, t is a given parameter, and solve for the optimal price and quality decisions, then, at optimality, we have $\Delta_q = \delta_m - \delta_o$ and $\Delta_p = p_m - p_o$ both decreasing in t.*

The immediate implication is that an increase in the stores' location differentiation will reduce the firm's quality and price differentiation at optimality. In other words, location differentiation can be treated as a substitute for the other two types of differentiation. Note that the substitution effect is also indicated in Corollary 1.

Note that we consider three types of differentiation – quality differentiation measured by $\Delta_q = \delta_m - \delta_o$, price differentiation defined by $\Delta_p = p_m - p_o$ and location differentiation quantified by t. An alternative way to understand their individual impacts is to analyze and compare models with and without one kind of differentiation. Since price and quality differentiation are always coupled together, we assume that either both of them exist or neither of them exists.

Proposition 4. *The individual effect of store differentiation is presented below:*

(a) In a model without location differentiation, if $F_o \geq \frac{2}{675c}$, the optimal strategy is to open the main store only, where $(\delta_m = \frac{1}{3c}, p_m = \frac{2}{9c})$ and $\Pi^m = \frac{1}{27c}$; otherwise, it is optimal to open both the main and outlet stores, where $(\delta_m = \frac{2}{5c}, p_m = \frac{7}{25c})$, $(\delta_o = \frac{1}{5c}, p_o = \frac{3}{25c})$ and $\Pi^{mo} = \frac{1}{25c} - F_o$.

(b) In a model without price and quality differentiation, it is optimal to open the main store only, where $(\delta_m = \frac{1}{3c}, p_m = \frac{2}{9c})$ *and* $\Pi^m = \frac{1}{27c}$.

The result above indicates that in the absence of price and quality differentiation, the outlet store would not be opened. As such, the addition of location differentiation on top of the price and quality differentiation can help mitigate the need to create differences in price and quality offerings (see Corollary 2).

2.4.3 The Effect of E-Commerce

In retailing, an irreversible trend is probably the rising of e-commerce, which imposes significant competition to physical stores. In response, many retailers, e.g., Macy's, are forced to either reduce the footage of the main stores or open/expand their outlet stores; see, e.g., Kieler (2016) and Wilson (2016). In view of these observations, in this subsection we consider the effect of e-commerce on the operations of the main and outlet stores. In particular, we will focus on the firm's store offering strategies and the contribution of the main store in the firm's total sales. Due to competition from e-commerce, some customers may decide to switch from the physical stores to online channels. We assume that, with probability $\gamma_m \in [0, 1]$, a customer who originally would buy from the main store will switch to buy from online channels, and with probability $\gamma_o \in [0, 1]$, a customer who originally would buy from the outlet store will switch to buy from online channels. We further assume that $\gamma_m \geq \gamma_o$, which implies that customers who originally buy from main store are more likely to switch to buy online than those who originally buy from outlet stores. A justification for this is as follows: Based on the demand characterization discussed in §3, customers who buy from main stores have higher valuation relative to those who buy from outlet stores. As such, it implies that higher valuation/income customers are more likely to buy online than lower valuation/income ones, due to, e.g., better access to internet and lower security risk, which is supported by both academic and trade literature, see, e.g., Horrigan (2008), Akman and Rehan (2014), and Smith (2015).

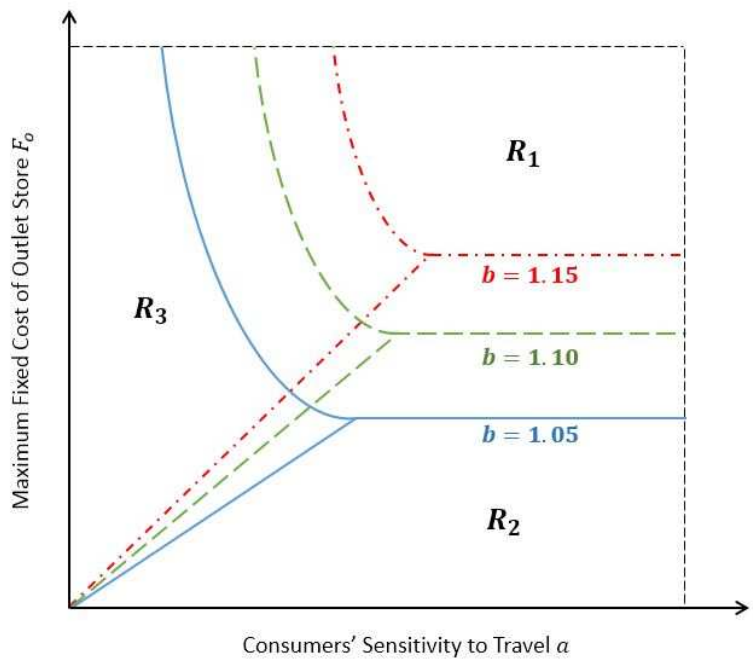
In the presence of e-commerce, the demand functions now become $Q'_m = (1-\gamma_m)Q_m$ and $Q'_o = (1-\gamma_o)Q_o$, where Q_m and Q_o are demand functions in the base model provided in §3. Since the potential population buying at the main and outlet stores is reduced (due to e-commerce), it is reasonable to assume that the fixed cost of operating the physical stores is also reduced proportionally to the reduction in the population size. So, the fixed costs are reduced to $F'_m = (1-\gamma_m)F_m$ and $F'_o = (1-\gamma_o)F_o$. Accordingly, the firm will set $(\delta_m, p_m, t, \delta_o, p_o)$ to maximize

$$
\begin{aligned}
\Pi &= (p_m - c_m)Q'_m + (p_o - c_o)Q'_o - (1-t)^2 F'_o - F'_m \\
&= (p_m - c_m)(1-\gamma_m)Q_m + (p_o - c_o)(1-\gamma_o)Q_o - (1-t)^2(1-\gamma_o)F_o - (1-\gamma_m)F_m \qquad (23) \\
&= (1-\gamma_m)[(p_m - c_m)Q_m + b(p_o - c_o)Q_o - b(1-t)^2 F_o - F_m],
\end{aligned}
$$

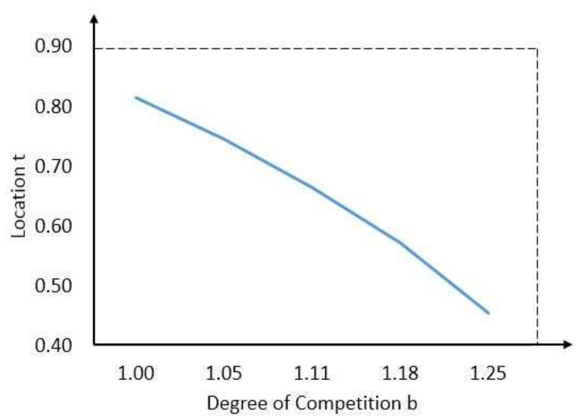
where $b = \frac{1-\gamma_o}{1-\gamma_m}$ and $b \geq 1$ since $\gamma_m \geq \gamma_o$. Following this formulation, we discuss a few points. First, the ratio b can be used to measure the degree of competition from e-commerce; that is, the higher b is, the stronger the competition from e-commerce. Letting $b = 1$ results in the base model without e-commerce. Second, it is clear that the firm's store offering and other decisions will depend on γ_m and γ_o only through their ratio b, rather than their individual values. Lastly, similar to the base model without e-commerce, F_m can be treated as a sunk cost and ignored in analysis. A similar analysis procedure is followed to solve for the firm's store offering strategy.

As shown in Appendix A (Part II), we were able to analytically derive the constrained optimal decisions and profit in case (m) and case (o). We use a numerical approach to derive the constrained optimal values in case (mo), followed by a comparison of the three cases.

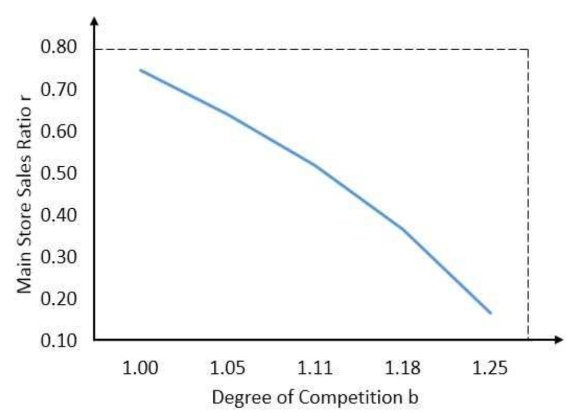
Figure 3(a), 3(b) and 3(c) presents the impact of e-commerce on i) the firm's optimal store offering strategy, ii) the optimal distance between the two stores, t, and iii) the contribution of the main store to the overall sales, measured by $r = \frac{Q'_m}{Q'_m + Q'_o}$, respectively. In the numerical study, we set $b \in [1, 1.25]$ to focus on the effect of e-commerce and fix $c = 0.1$. In Figures 3(b) and 3(c), we further fix $F_o = 0.05$ and $a = 0.2$, under which both the main and outlet stores are operational with positive sales.



(a) on the store offering strategy



(b) on the store distance t



(c) on contribution of the main store

Figure 2.3: Impact of Competition from E-Commerce

There are a number of interesting observations. First, Figure 3(a) indicates that the stronger the competition from e-commerce, the more likely that the firm will forgo the strategy of using the main store only, that is, size of region R_1 is reduced. Instead, the firm becomes more likely to add the outlet store in the same location of the main store, that is, size of region R_2 increases. Second, according to Figures 3(b) and 3(c), as the competition from the e-commerce increases, if both the main and outlet stores are active, they would be located closer, that is, t is lower, and the sales contribution of the main store will decrease. These observations are consistent with the anecdotal evidence of increasing sales of outlet stores at Macy's, see, e.g., Kieler (2016).

2.5. Model Extension: Heterogeneous Travel Sensitivity

So far we have assumed that all consumers are homogeneous in their sensitivity to travel. However, due to their heterogeneity in valuation of the product, they may have different sensitivity to travel as well. For example, customers' willingness to pay for a product might be contingent on their income level. Oftentimes, higher willingness to pay is associated with higher income levels, which translates into higher disutility for time to travel. More specifically, we assume that customers with a lower valuation of the product (i.e., $v \in [0, 0.5]$) will have a travel sensitivity coefficient a_1 and customers with a higher valuation of the product (i.e., $v \in (0.5, 1]$) will have a travel sensitivity coefficient a_2, where $0 \leq a_1 \leq a_2 \leq 1$. Note that our assumption that $a_1 \leq a_2$ is supported by both economic theory (see, e.g., Hill (1985) that higher income consumers tend to have higher costs of personal time) and empirical studies (see, e.g., Ngwe (2014) that consumers' travel sensitivity is positively correlated with their taste for quality).[1] When $a_1 = a_2$, the model converges to the base model studied in the previous section.

Given the heterogeneity in travel sensitivity, we first characterize the demand function for the main and outlet stores based on consumers' utility functions. If a customer buys from the main store, its utility is the same as that in the base model where $U_m = \delta_m v - p_m$, where $v \sim U[0,1]$. However, depending on where the customers' valuation of the product, their utility from buying at the outlet store can be written as

$$U_o = \begin{cases} \delta_o v - a_1 t - p_o & \text{if } \ v \in [0, 0.5] \\ \delta_o v - a_2 t - p_o & \text{if } \ v \in (0.5, 1] \end{cases}$$

According to the utility functions above, we derive demand for the main and outlet stores in each of the two customer segments – segment 1 refers to $v \in [0, 0.5]$ and segment 2 refers to the remaining valuations. So, we use notation (Q_m^1, Q_m^2) to represent customers' demand for the main store from segments 1 and 2, respectively, and (Q_o^1, Q_o^2) for the outlet store from the two segments. If we characterize these four demand quantities based on whether they are zero (i.e., no sales) or non-zero (i.e., positive sales), there are sixteen possible different combinations. To simplify the presentation, we first focus on the demand at the main store, which gives us four combinations presented in the following table:

[1]Note that we have $v = 0.5$ as the cutoff for high and low valuation customers. Indeed, our model can be extended for any cutoff value in [0,1].

Table 2.4: Demand Scenarios in the Main Store

	$Q_m^2 = 0$	$Q_m^2 > 0$
$Q_m^1 = 0$	No sales at the main store.	Only customers from segment 2 will buy from the main store.
$Q_m^1 > 0$	Not feasible.	Positive sales at the main store from both segments.

Our analysis in Appendix B indicates that it is never possible to set decisions within their boundaries such that some customers from segment 1 will buy from the main store but no customers in segment 2 would do so. That is, the case with $(Q_m^1 > 0, Q_m^2 = 0)$ is not feasible. In our subsequent equilibrium analysis of the firm's decisions, it can be verified that it is never optimal to have either no sales at all or positive sales from both segments at the main store. That is, the cases with $(Q_m^1 = 0, Q_m^2 = 0)$ and $(Q_m^1 > 0, Q_m^2 > 0)$ are not optimal. Hence, the only combination remaining is the one in the top right corner, $(Q_m^1 = 0, Q_m^2 > 0)$, when no customers with low valuation (in segment 1) would buy in the main store, but some customers with high valuation (in segment 2) would do so. To be more specific, we present the demand functions in this scenario in Table 5 below.

Table 2.5: Demand Functions when $(Q_1^m = 0, Q_2^m > 0)$

	Q_m^1	Q_m^2	Q_o^1	Q_o^2
Case (1)	0	$1 - \frac{p_m}{\delta_m}$	0	0
Case (2)	0	$1 - \frac{p_m}{\delta_m}$	$\frac{1}{2} - \frac{p_o + a_1 t}{\delta_o}$	0
Case (3)	0	$1 - \frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}$	0	$\frac{p_m - p_o - a_2 t}{\delta_m - \delta_o} - \frac{p_o + a_2 t}{\delta_o}$
Case (4)	0	$1 - \frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}$	$\frac{1}{2} - \frac{p_o + a_1 t}{\delta_o}$	$\frac{p_m - p_o - a_2 t}{\delta_m - \delta_o} - \frac{p_o + a_2 t}{\delta_o}$

Note from Appendix B that the conditions for $(Q_1^m = 0, Q_2^m > 0)$ are $p_o \geq p_m - a_2 t + \delta_o - \delta_m$ (to ensure $Q_2^m > 0$) and $p_o \leq p_m - a_1 t + \frac{\delta_o}{2} - \frac{\delta_m}{2}$ if $p_m \leq \frac{\delta_m}{2}$ (to ensure $Q_1^m = 0$). Under these conditions, for any given set of decision variables, we can characterize demand functions which fit into one and only one of the corresponding four cases presented in Table 5. This is an equivalent way to provide the conditions for each of the four demand cases to occur.

- Case (1): $p_m \geq \frac{\delta_m}{2}$, $p_o \geq \frac{\delta_o}{2} - a_1 t$ and $p_o \geq \frac{\delta_o p_m}{\delta_m} - a_2 t$;
- Case (2): $p_m \leq \frac{\delta_m}{2}$, $p_o \leq p_m - a_1 t + \frac{\delta_o}{2} - \frac{\delta_m}{2}$ and $p_o \geq \frac{\delta_o p_m}{\delta_m} - a_2 t$;
- Case (2): $p_m \geq \frac{\delta_m}{2}$, $p_o \leq \frac{\delta_o}{2} - a_1 t$ and $p_o \geq \frac{\delta_o p_m}{\delta_m} - a_2 t$;

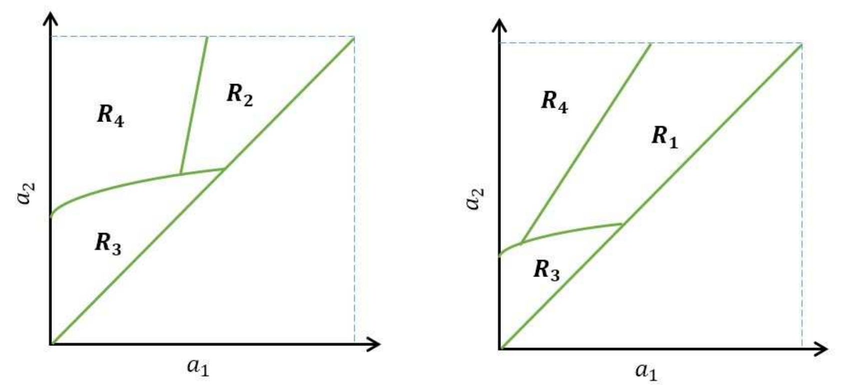
- Case (3): $p_m \geq \frac{\delta_m}{2}$, $p_o \geq \frac{\delta_o}{2} - a_1 t$ and $p_m - a_2 t + \delta_o - \delta_m \leq p_o \leq \frac{\delta_o p_m}{\delta_m} - a_2 t$;
- Case (4): $p_m \leq \frac{\delta_m}{2}$, $p_o \leq p_m - a_1 t + \frac{\delta_o}{2} - \frac{\delta_m}{2}$ and $p_m - a_2 t + \delta_o - \delta_m \leq p_o \leq \frac{\delta_o p_m}{\delta_m} - a_2 t$; and
- Case (4): $p_m \geq \frac{\delta_m}{2}$, $p_o \leq \frac{\delta_o}{2} - a_1 t$ and $p_m - a_2 t + \delta_o - \delta_m \leq p_o \leq \frac{\delta_o p_m}{\delta_m} - a_2 t$.

Given the demand functions, the firm can set the decisions on two pricing (p_m and p_o), two quality (δ_m and δ_o) and one location (t) to maximize

$$\Pi^{mo} = (p_m - c_m)(Q_m^1 + Q_m^2) + (p_o - c_o)(Q_o^1 + Q_o^2) - F_o(1-t)^2.$$

In order to find the global optimal decisions for the firm, we need to analyze the four cases presented in Table 5 and compare the constrained optimal solution in each case. The optimal decisions include choosing $\delta_m, \delta_o, p_m, p_o$, and t. The constrained optimal solution in case (1) can be solved analytically and is provided in Appendix B. We can also obtain the unconstrained optimal solution in cases (2)– (3) analytically in Appendix B. In case (4), we were able to characterize the unconstrained optimal solution in closed form for a given δ_o and parametric space (c, F_o, a_1, a_2) in Appendix B. However, given the complication in the firm's profit function in δ_o, it is analytically challenging to derive the unconstrained optimal δ_o in this case. Hence, we resort to numerical analysis to obtain the constrained optimal δ_o and the other corresponding decisions and profit. Without loss of generality, we have used $c = 0.1$ throughout the numerical study here. We will focus on the other three parameters F_o, a_1 and a_2. In particular, our numerical study indicates that the effect of travel sensitivity heterogeneity depends on whether F_o is below or above a threshold value, $F'_o = \frac{2}{675c}$, which also appears in the base model in Figure 2. Hence, the two values we take for F_o are $(\frac{19}{20})(\frac{2}{675c}) < F'_o$ and $(4)(\frac{2}{675c}) > F'_o$ which represent cases of low and high fixed cost of the outlet store (compared to the threshold F'_o), respectively.

Similar to the case with homogeneous travel sensitivity, we can plot the firm's optimal store strategy (whether to operate both the main and outlet stores) in the (a_1, a_2) plane. In our numerical examples, the travel sensitivity a_1 and a_2 range from 0 to 1, where $a_2 \geq a_1$. So we only need to examine the upper left triangle. Moreover, due to different strategies derived under a low or a high fixed cost for the outlet store, we plot the case with a low fixed cost in Figure 4(a) and the case with a high fixed cost in Figure 4(b).



(a) Low Fixed Cost: $F_o = \frac{19}{20}(\frac{2}{675c})$

(b) High Fixed Cost: $F_o = 4(\frac{2}{675c})$

Figure 2.4: Equilibrium Regions with Heterogeneous Travel Sensitivity

According to the equilibrium regions presented in Figure 4, we plot the corresponding demand distribution/sales of both stores in each region in Figure 5 to facilitate the discussion on the impact of heterogeneous travel sensitivity.

According to Figures 4 and 5, similar to Figure 2 in the homogeneous travel sensitivity case, when sensitivity becomes heterogeneous, we still have region R_1 where only the main store is effective with positive sales, region R_2 where both stores have positive sales and are operated at the same location, and region R_3 where both stores have positive sales but are maintained with a distance. In both regions R_2 and R_3, the customers who decide to buy are all high value customers and located in segment 2 with $v \geq 0.5$. This observation is also consistent with that in the homogeneous travel sensitivity case. In addition to these three regions (R_1,

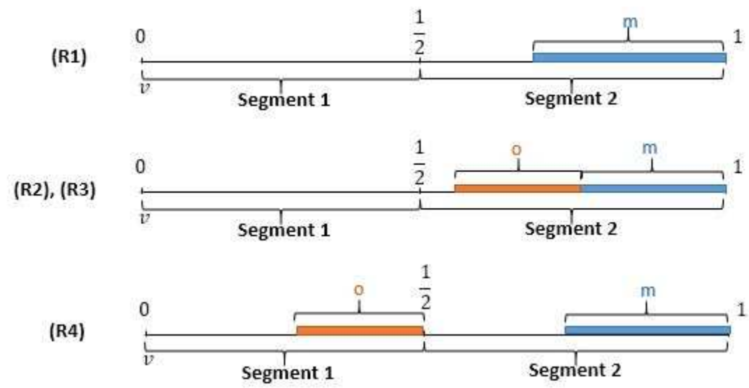


Figure 2.5: Demand Distributions in Equilibrium Regions

R_2 and R_3), the heterogeneity in travel sensitivity results in a new region R_4 as a potential equilibrium offering strategy. In this region, both main and outlet stores are active and operated at different locations. Moreover, the low valuation customers in segment 1 will buy only from the outlet store and the high valuation customers in segment 2 will buy only from the main store. This new region leads to the following conclusion:

Observation 1. *The firm will use the main store (resp., the outlet store) to target the high (resp., low) valuation customers when the degree of travel sensitivity heterogeneity (measured by $\triangle_a = a_2 - a_1$) is sufficiently high.*

An immediate implication is that heterogeneous travel sensitivity helps the firm to use the two stores to differentiate its customer segments better. When a_1 and a_2 are close to each other, customers in both segments are relatively equally sensitive to travel and the model is reduced to the base case with homogeneous travel sensitivity. Then, the firm is more willing to set its pricing and quality decisions so as to target only the high valuation customers (since targeting also the low valuation ones implies that the price and/or quality at the outlet store has to be low and the pricing and/or quality decisions between the two stores have to be highly differentiated). However, when a_2 is significantly higher than a_1, the two segments of customers are already highly differentiated in their travel sensitivity, which makes it much easier for the firm to set decisions to direct each of the two customer segments into its corresponding stores.

Finally, comparison of the high and low fixed cost cases in Figure 4(a) and (b) indicates that it is never optimal to operate the main store only (respectively, the two stores at the same location) if the fixed cost of having an outlet store is low (respectively, high). This difference matches well with the observation in Figure 2 which is a special case of Figure 4 when $a_1 = a_2$.

In the rest of this section, we will examine how travel heterogeneity affects the firm's profitability. The numerical result is presented in Figure 6, and we fix $F_o = (\frac{19}{20})(\frac{2}{675c})$, a value also used in Figure 4.

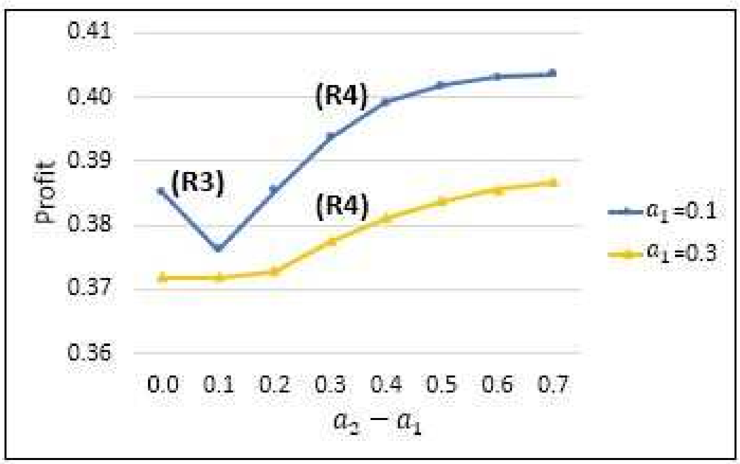


Figure 2.6: Impact of Travel Heterogeneity on the Firm's Profit

Figure 6 indicates that the firm's profit generally increases in the degree of travel heterogeneity (corresponding to R_4), except when both a_1 and a_2 are low in which case the firm's profit decreases as the degree of travel heterogeneity increases (corresponding to R_3). In R_4, the firm's optimal store offering strategy is to segment the consumer groups. Since higher degree of travel heterogeneity implies higher differentiation between the two groups of customers, the firm can use this to better segment the customers in setting the prices and qualities of its products sold in the two stores, and hence gain from it. In R_3, when both a_1 and a_2 are small, the firm's optimal store offering strategy is to offer both stores at a distance and set prices and qualities in the two stores to compete for the high value customers. An increase in the degree of travel heterogeneity will result in a higher travel sensitivity of high-valuation consumers, which actually undermines the firm's profit since now both stores are targeting the high-valuation customers (as in the base case). Then, an increase in the travel sensitivity will make travel more costly which will discourage the customers to buy from the outlet store (but not necessarily encourage them to buy from the main store). So, the firm's profit will be reduced as travel heterogeneity increases.

2.6. Conclusion

Outlet stores have existed for a long time and have increasingly become an integral part of a firm's retail distribution strategy. The role of outlet stores has evolved from traditional disposal of surplus inventory as an operational practice to nowadays being considered as a distribution channel strategy in order to achieve customer segmentation, market differentiation and wider coverage. It is generally accepted that the main and outlet stores offer differentiated products in the sense that the main store offers a product with a higher quality but at a higher price while the product in the outlet store usually has a lower quality and is also sold at a lower price. In view of this, on one hand, the outlet store can help the firm better segment its heterogeneous customer group and achieve higher market coverage since customers whose valuation might be too low to buy from the main store can now buy from the outlet store. But, the outlet store may cannibalize sales in the main store where prices are usually higher and hence undermine the firm's overall profitability. In addition to product quality and price differentiation between the main and outlet stores, we introduce a third type of differentiation between the two stores, i.e., the location differentiation. Since the outlet store is usually located far away from the main store (where the consumer base is) in order to save cost, there is a disutility for consumers who travel to the outlet store. So, customers need to consider the quality and price of the product and the travel cost when they decide whether and where to buy. In this paper, we formulate a stylized economic model to gain understanding of the trade-offs faced by the firm and the interplay among the three types of differentiation between the main and outlet stores.

Our analysis of the base model where customers are heterogeneous in product valuation but are equally sensitive to travel to the outlet store, indicates that the outlet store is more likely to be introduced when (1) the fixed cost to open the outlet store or the base unit production cost is lower; or (2) the customers are more willing to travel to the outlet store. Moreover, introduction of the outlet store would actually encourage the firm to improve the quality of the product sold in the main store so as to differentiate relative to the product sold in the outlet store. These results have implications for both the firm and the customers. The firm should seriously consider the outlet store strategy if the customers are not very sensitive to travel. Indeed, the higher the (fixed or unit production) costs associated with the outlet store are, the further the outlet store should be located away from the main store. From the perspective of customers, they should be less concerned about the possible deterioration in the quality of the product sold in the main store. Our analysis also demonstrates that location differentiation can be considered as a substitute to product quality and price differentiation. We also consider the effect of competition from e-commerce and show that it makes the firm more likely to introduce the outlet store and even bring it closer to the main store. The resulting effect of reduced location differentiation is to enhance quality and price differentiation in offerings between the two stores.

In an extension when customers are heterogeneous in both product valuation and in travel sensitivity, we show that it becomes more likely for the firm to use the segmentation strategy where high valuation customers buy from the main store and the low valuation customers buy from the outlet stores. A direct implication from it is that the more different customers are, it becomes easier for the firm to reduce the cannibalization effect and could lead to a higher profit.

Finally, note that we have made a number of simplifying assumptions for model tractability purpose. One of the main assumptions is that we considered a monopoly firm in the market and competition from e-commerce was formulated in an exogenous manner. There might be multiple firms competing in the market who would consider the option of opening outlet stores. Introducing horizontal competition is an interesting extension which would lead to different decisions and insights in terms of whether or not to open an outlet store.

2.7. Bibliography

[1] Akman I, Rehan M. Online purchase behaviour among professionals: a socio-demographic perspective for Turkey. *Economic Research-Ekonomska Istraivanja*, 2014, 27(1): 689-699.

[2] Anderson E T, Dana Jr J D. When is price discrimination profitable? *Management Science*, 2009, 55(6): 980-989.

[3] Banker R D, Khosla I, Sinha K K. Quality and competition. *Management science*, 1998, 44(9): 1179-1192.

[4] Bell D R, Wang Y, Padmanabhan V. The effect of partial forward integration on retailer behavior: An explanation for co-located stores. Working paper, University of Pennsylvania, 2003.

[5] Bernstein F, Federgruen A. A general equilibrium model for industries with price and service competition. *Operations research*, 2004, 52(6): 868-886.

[6] Brandculture. 2010. https://www.brandculture.com/when-is-a-coach-bag-not-really-a-coach-bag-when-its-from-the-coach-factory-store/

[7] Cao J, So K C, Yin S. Impact of an online-to-store channel on demand allocation, pricing and profitability. *European Journal of Operational Research*, 2016, 248(1): 234-245.

[8] Chiang W K, Chhajed D, Hess J D. Direct marketing, indirect profits: A strategic analysis of dual-channel supply-chain design. *Management science*, 2003, 49(1): 1-20.

[9] Coughlan A, Soberman D A. A survey of outlet mall retailing: Past, present, and future. INSEAD, 2004.

[10] Coughlan A T, Soberman D A. Strategic segmentation using outlet malls. *International Journal of Research in Marketing*, 2005, 22(1): 61-86.

[11] Deneckere R J, Preston McAfee R. Damaged goods. *Journal of Economics & Management Strategy*, 1996, 5(2): 149-174.

[12] Hill M. Investments of time in houses and durables. Time, Goods and Well-Being, University of Michigan, Ann Arbor, 1985: 205-243.

[13] Horrigan B J. Online Shopping. 2008. http://www.pewinternet.org/2008/02/13/part-3-low-income-internet-users-and-online-shopping/.

[14] Humphers L. States of the outlet industry. *Value Retail News*, 2015.

[15] Iyer G. Coordinating channels under price and nonprice competition. *Marketing Science*, 1998, 17(4): 338-355.

[16] Kapner S. Retailers Lines Blur on Outlet Stores. 2014. http://www.wsj.com/articles/retailers-lines-blur-on-outlet-stores-1413159747.

[17] Kekre S, Srinivasan K. Broader product line: a necessity to achieve success? *Management science*, 1990, 36(10): 1216-1232.

[18] Kieler A. Macys Opening 16 Additional Discount Backstage Stores This Year. 2016. https://consumerist.com/2016/02/23/macys-opening-16-additional-off-di.

[19] Kok A G, Fisher M L. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 2007, 55(6): 1001-1021.

[20] Kumar V, Venkatesan R. Who are the multichannel shoppers and how do they perform?: Correlates of multichannel shopping behavior. *Journal of Interactive marketing*, 2005, 19(2): 44-62.

[21] Lieber C. Buyer Beware: What You're Actually Getting at Outlet Stores. 2014. http://www.racked.com/2014/10/8/7573957/outlet-mall-stores.

[22] Maheshwari S. Consumers May Not Know Theyre Getting Lower-Quality Clothes At Outlet Stores. 2014.

[23] Manez J A, Waterson M. Multiproduct firms and product differentiation: a survey. Working Paper, University of Warwick, 2001.

[24] Matsubayashi N. Price and quality competition: The effect of differentiation and vertical integration. *European Journal of Operational Research*, 2007, 180(2): 907-921.

[25] Myers J B, Pickersgill A D, Van Metre E S. Steering customers to the right channels. *McKinsey quarterly*, 2004, 4: 36-47.

[26] Moorthy K S. Strategic decentralization in channels. *Marketing Science*, 1988, 7(4): 335-355.

[27] Mussa M, Rosen S. Monopoly and product quality. *Journal of Economic theory*, 1978, 18(2): 301-317.

[28] Neslin S A, Shankar V. Key issues in multichannel customer management: current knowledge and future directions. *Journal of interactive marketing*, 2009, 23(1): 70-81.

[29] Ngwe D. Why Outlet Stores Exist: Averting Cannibalization in Product Line Extensions. Working paper, Harvard University, 2014.

[30] Qian Y, Anderson E, Simester D. Multichannel spillovers from a factory store. Working Paper 19176, National Bureau of Economic Research, 2013. http://www.nber.org/papers/w19176.

[31] Rocha C. The Truth About TJ Maxx and Marshalls. 2010. http://www.pocketyourdollars.com/2010/06/the-truth-about-tj-maxx-and-marshalls/.

[32] Ryzin G, Mahajan S. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science*, 1999, 45(11): 1496-1509.

[33] Smith C. The surprising facts about who shops online and on mobile. 2015. http://www.businessinsider.com/the-surprising-demographics-of-who-shops-online-and-on-mobile-2014-6/

[34] Tsay A A, Agrawal N. Channel conflict and coordination in the ecommerce age. *Production and Operations Management*, 2004, 13(1): 93-110.

[35] Villas-Boas J M. Product line design for a distribution channel. *Marketing Science*, 1998, 17(2): 156-169.

[36] Wilson M. Department store giant continues to expand outlet store formats. 2016. http://www.chainstoreage.com/article/department-store-giant-continues-expand-outlet-store-formats.

[37] Zhang T. Outlet vs. Retail: Risk of Cannibalization? 2016. http://www.overtcollusion.com/pricing/2016/2/22/outletvsretailriskofcannibalization.html.

2.8. Appendix

Appendix A (Part I): Base Model with Homogeneous Travel Sensitivity

Proof of Lemma 1: In this proof, we solve for the constrained optimal decisions and profit in each of the three cases (m), (o) and (mo). We start with case (m) first.

Case (m): In this case, only the main store has positive sales and the condition that enables this is $p_o \geq \frac{p_m \delta_o}{\delta_m} - at$. Given this condition, the firm sets δ_m and p_m to maximize

$$\Pi^{mo} = (p_m - c\delta_m^2)(1 - \frac{p_m}{\delta_m}), \quad subject \ to \ \ \delta_m \geq 0 \ \ and \ \ 0 \leq p_m \leq \delta_m.$$

It is straightforward to verify that the firm's profit function is concave in p_m for a given δ_m and the unconstrained optimal selling price in the main store is

$$p_m = \frac{\delta_m(1 + c\delta_m)}{2} \geq 0.$$

Note that this unconstrained p_m is also less than δ_m if $c\delta_m \leq 1$. If $c\delta_m > 1$, it is easy to verify that there does not exist a proper value of p_m so that the firm can have both non-negative profit margin, measured by $(p_m - c\delta_m^2)$, and non-negative sales/demand, measured by $(1 - \frac{p_m}{\delta_m})$. In other words, it is never optimal for the firm to set $c\delta_m > 1$. Substituting $p_m = \frac{\delta_m(1+c\delta_m)}{2}$ into the firm's profit function and simplifying yields $\Pi^{mo} = \frac{1}{4}\delta_m(1 - c\delta_m)^2$. It is fairly easy to show that there exists a global optimal product quality

$$\delta_m = \frac{1}{3c} \in [0, \frac{1}{c}]$$

that maximizes the firm's profit. The resulting optimal profit in case (m) is

$$\Pi^{mo} = \frac{1}{27c}.$$

In terms of the decisions associated with the outlet store (t, δ_o, p_o), they need to satisfy the condition stated earlier, i.e., $p_o \geq \frac{p_m \delta_o}{\delta_m} - at$.

Case (o): In this case, only the outlet store has positive sales and the condition that enables this is $\delta_m - p_m \leq \delta_o - p_o - at$. Given this condition, the firm sets (t, δ_o, p_o) to maximize

$$\Pi^{mo} = (p_o - c\delta_o^2)(1 - \frac{p_o + at}{\delta_o}) - (1-t)^2 F_o, \;\; subject\;\; to\;\; 0 \leq t \leq 1, \;\; \delta_o \geq 0 \;\; and \;\; 0 \leq p_o \leq \delta_o.$$

Similar to case (m), the firm will never set δ_o such that $c\delta_o > 1$ since it makes it impossible to obtain a positive total profit. Using a similar approach as in case (m), for any given (t, δ_o), the firm's profit function is concave in the selling price p_o and the unconstrained value is

$$p_o = \frac{\delta_o}{2}(1 + c\delta_o) - \frac{at}{2} \leq \delta_o.$$

Note that this value of p_o would be non-negative if $\delta_o \geq at$, which will be satisfied, since, if not, the outlet store will not have positive sales/demand at all.

Substituting $p_o = \frac{\delta_o}{2}(1 + c\delta_o) - \frac{at}{2}$ into the firm's profit function and simplifying yields $\Pi^{mo} = \frac{1}{4}c^2\delta_o^3 - \frac{1}{2}c\delta_o^2 + (\frac{1}{4} + \frac{1}{2}act)\delta_o + \frac{1}{4}\frac{a^2t^2}{\delta_o} - (1-t)^2 F_o$. With some algebra, it can be shown that there exists a unique constrained global optimal value of δ_o, where

$$\delta_o = \frac{1 + \sqrt{1 + 12act}}{6c}.$$

Substituting this value back into the profit function and simplifying yields a function of decision variable t only,

$$\Pi^{mo} = \frac{(1 - 12act + \sqrt{1 + 12act})^2}{54c(1 + \sqrt{1 + 12act})} - (1 - t)^2 F_o$$

which has two stationary points : $t = \frac{a^3c - 2aF_o + 6F_o^2 \pm \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}$. The general approach is to calculate the firm's profit at all stationary points and at the extreme points of the corresponding decision set and compare these profits to derive the best profit, which will also identify the optimal solution of the decisions. With some algebra, we can rule out $t = \frac{a^3c - 2aF_o + 6F_o^2 - \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}$ since the first-order condition of the profit with respect to t at this point is positive which implies that the firm's profit is still increasing as t increases. The other stationary point,

$$t = \frac{a^3c - 2aF_o + 6F_o^2 + \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}$$

is possible to be optimal since the first-order condition at this point is negative. With some algebra, we can show that this value of t is always less than 1. Further, it is non-negative if and only if $F_o \geq \frac{a}{6}$.

Hence, for $F_o \leq \frac{a}{6}$, we need compare the firm's profit at $t = 0$ and at $t = 1$, which results with $t = 0$ being optimal. Accordingly, $(\delta_o = \frac{1}{3c}, p_o = \frac{2}{9c})$ and $\Pi^{mo} = \frac{1}{27c} - F_o$.

For $F_o \geq \frac{a}{6}$, we need to compare the firm's profit at $t = \frac{a^3c - 2aF_o + 6F_o^2 + \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}$ (i.e., the interior solution) and at the two extreme points $t = 0$ and $t = 1$, which results with the interior solution being optimal. Accordingly, we have $(\delta_o = \frac{1 + \sqrt{1 + 12act}}{6c}, p_o = \frac{1 - 3act + \sqrt{1 + 12act}}{9c})$ and $\Pi^{mo} = \frac{(1 - 12act + \sqrt{1 + 12act})^2}{54c(1 + \sqrt{1 + 12act})} - (1 - t)^2 F_o$.

In summary, the optimal decisions and profit in case (o) are as follows:

- If $F_o \leq \frac{a}{6}$: $(t = 0, \delta_o = \frac{1}{3c}, p_o = \frac{2}{9c})$ and $\Pi^{mo} = \frac{1}{27c} - F_o$.

- Otherwise:
$(t = \frac{a^3c - 2aF_o + 6F_o^2 + \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}, \delta_o = \frac{1+\sqrt{1+12act}}{6c}, p_o = \frac{1-3act+\sqrt{1+12act}}{9c})$ and $\Pi^{mo} = \frac{(1-12act+\sqrt{1+12act})^2}{54c(1+\sqrt{1+12act})} - (1-t)^2F_o$.

Case (mo): In this case, both the main and outlet stores have positive sales, and it is the most complicated case to analyze. The firm sets (δ_m, p_m) and (t, δ_o, p_o) to maximize

$$\Pi^{mo} = (p_m - c\delta_m^2)(1 - \frac{p_m - p_o - at}{\delta_m - \delta_o}) + (p_o - c\delta_o^2)(\frac{p_m - p_o - at}{\delta_m - \delta_o} - \frac{p_o + at}{\delta_o}) - (1-t)^2F_o,$$

subject to $0 \leq \delta_o \leq \delta_m$, $0 \leq p_o \leq \delta_o$, $0 \leq p_m \leq \delta_m$, $0 \leq t \leq 1$ and $p_m + \delta_o - \delta_m - at \leq p_o \leq \frac{p_m\delta_o}{\delta_m} - at$. Given many decisions and constraints involved, here is our solution approach. First, we will solve the first-order-conditions (or equations) to derive all possible stationary points. And then, we will calculate the firm's profit at stationary points (given they are within the constraints) and at the boundary points of the parametric space and compare these profits to derive the best value. According to the first-order conditions, we were able to derive eight possible stationary points. Seven of these eight points either violate at least one of the constraints or have the Hessian matrix that is not negative definite (which is a necessary condition for the stationary point to be at least a local optima). For space consideration, we will not list these seven stationary points here. The only stationary point that can possibly be locally and globally optimal is as follows:

$$t = \hat{t} = \frac{-9a^3c + 2aF_o + 10F_o^2 - 3\sqrt{9a^6c^2 - 4a^4cF_o + a^2F_o^2 - 20a^3cF_o^2}}{10F_o^2},$$

$$\delta_m = \frac{B}{5c}, \quad p_m = \frac{-20A + 7B}{50c}, \quad \delta_o = \frac{B}{10c}, \quad p_o = \frac{3(-10A + B)}{50c},$$

where $A = 20act$ and $B = 1 + \sqrt{1 - 20A}$. We next need to check whether this point satisfies the constraints of the problem. It is easy to show that $0 \leq \delta_o \leq \delta_m$, $0 \leq p_o \leq \delta_o$ and

$0 \le p_m \le \delta_m$. The two important sets of constrains that need special attention are $0 \le t \le 1$ and $p_m + \delta_o - \delta_m - at \le p_o \le \frac{p_m \delta_o}{\delta_m} - at$. Let us first look at the second set of constraints, $p_m + \delta_o - \delta_m - at \le p_o \le \frac{p_m \delta_o}{\delta_m} - at$. Note that if any of the two constraints in this set is violated after substituting the values of the stationary point in, then we can conclude that the firm makes a negative profit in either the main store or the outlet store since the corresponding demand is negative. So the total profit from both the main and the outlet stores under this stationary point will be dominated by either case (m) or case (o) (which are the two possible boundary solutions of the problem). Given this observation, we do not need to specifically check this set of constraints since they will be taken care of when we compare the profit under this stationary point with the profits under boundary solutions.

In terms of the first set of constraints, $0 \le t \le 1$, with some algebra, we can show that $\hat{t} < 1$ always holds true, and that $\hat{t} \ge 0$ when $F_o \ge \frac{a}{10}$.

Hence, for $F_o \le \frac{a}{10}$, we need to compare the firm's profit at $t = 0$ (where the other decisions can be calculated based on $\delta_m = \frac{B}{5c}, p_m = \frac{-20A+7B}{50c}, \delta_o = \frac{B}{10c}, p_o = \frac{3(-10A+B)}{50c}$) and at the boundary points of the problem. Note that it is clear that the boundary solutions at either $\delta_o = 0$, $\delta_o = \delta_m$, $p_o = 0$, $p_o = \delta_o$, $p_m = 0$ or $p_m = \delta_m$ can never be optimal for the firm. So the effective boundary solutions here are $t = 1$, or $p_o = \frac{p_m \delta_o}{\delta_m} - at$ (i.e., case (m)) or $p_m + \delta_o - \delta_m - at = p_o$ (i.e., case (o)). It turns out that $t = 0$ dominates all of these boundary solutions if $F_o \le \frac{2}{675c}$; otherwise, case (m) dominates all other possibilities. Hence, we conclude that, if $F_o \le \min(\frac{a}{10}, \frac{2}{675c})$, the firm's optimal decisions are $(\delta_m = \frac{2}{5c}, p_m = \frac{7}{25c})$, $(t = 0, \delta_o = \frac{1}{5c}, p_o = \frac{3}{25c})$ and the corresponding profit is $\Pi^{mo} = \frac{1}{25c} - F_o$; otherwise, case (m) dominates.

For $F_o \ge \frac{a}{10}$, we have the stationary point as the interior solution, and we need to compare the firm's profit at this interior stationary point with those at the boundary points of the problem, which are $t = 0$, or $t = 1$, or $p_o = \frac{p_m \delta_o}{\delta_m} - at$ (i.e., case (m)) or $p_m + \delta_o - \delta_m - at = p_o$ (i.e., case (o)). It turns out that the interior stationary point dominates all the boundary solutions if $\frac{a}{10} \le F_o \le \bar{F}_o$, where $\bar{F}_o = 1$ if $a \le \frac{1}{36c}$ and $\bar{F}_o = \frac{9(-3a^2c+\sqrt{15}\sqrt{-a^4c^2+36a^5c^3})}{8(-2+45ac)})$ otherwise. The corresponding profit is $\Pi^{mo} = \frac{200A^2-10A+B}{25Bc} - (1-t)^2 F_o$; otherwise, case (m) dominates the rest of the possibilities.

Given the analysis above, we can summarize the optimal decisions and profit as follows:

- If $F_o \le \min(\frac{a}{10}, \frac{2}{675c})$, the optimal decisions are $(\delta_m = \frac{2}{5c}, p_m = \frac{7}{25c})$ for the main store and $(t = 0, \delta_o = \frac{1}{5c}, p_o = \frac{3}{25c})$ for the outlet store, and the optimal profit is $\Pi^{mo} = \frac{1}{25c} - F_o$.
- If $\frac{a}{10} \le F_o \le \bar{F}_o$, the optimal decisions are $(\delta_m = \frac{B}{5c}, p_m = \frac{-20A+7B}{50c})$ for the main store and $(t = \hat{t} = \frac{-9a^3c+2aF_o+10F_o^2-3\sqrt{9a^6c^2-4a^4cF_o+a^2F_o^2-20a^3cF_o^2}}{10F_o^2}, \delta_o = \frac{B}{10c}, p_o = \frac{3(-10A+B)}{50c})$ for the outlet store, and the optimal profit is $\Pi^{mo} = \frac{200A^2-10A+B}{25Bc} - (1-t)^2 F_o$, where $A = 20act$ and $B = 1 + \sqrt{1-20A}$.
- In all other cases, case (m) dominates resulting with the highest profit for the firm.

The above analysis completes the proof of Lemma 1.

Proof of Propositions 1 and 2: The firm's optimal store offering strategy and the corresponding optimal decisions and profit can be derived by comparing the firm's profits in the three cases analyzed in the proof of Lemma 1 – cases (m), (mo) and (o). The comparison will lead to the following outcome.

- If $0 \le F_o \le \min(\frac{a}{10}, \frac{2}{675c})$ (Region R.2), we show that case (mo) dominates the other two cases and hence becomes the firm's best strategy to offer both the main and outlet stores. Further, the two stores will be located at the same site due to low fixed cost of opening an outlet store so that the firm can afford to have it next to the main store. Accordingly, the optimal decisions are: for the main store, $(\delta_m = \frac{2}{5c}, p_m = \frac{7}{25c})$, which leads to sales of the main store being $Q_m = 1 - \frac{p_m - p_o - at}{\delta_m - \delta_o} = \frac{1}{5}$; for the outlet store, $(t = 0, \delta_o = \frac{1}{5c}, p_o = \frac{3}{25c})$, which leads to the sales of the outlet store being $Q_o = \frac{p_m - p_o - at}{\delta_m - \delta_o} - \frac{p_o + at}{\delta_o} = \frac{1}{5}$. The firm's optimal profit is $\Pi^{mo} = \frac{1}{25c} - F_o$, which is composed by its profit in the main store $\Pi_m^{mo} = (p_m - \delta_m^2 c)Q_m = \frac{3}{125c}$ and that in the outlet store $\Pi_o^{mo} = (p_o - \delta_o^2 c)Q_o - (1-t)^2 F_o = \frac{2}{125c} - F_o$.

- If $\frac{a}{10} \leq F_o \leq \bar{F}_o$ (Region R.3), where $\bar{F}_o = 1$ if $a \leq \frac{1}{36c}$ and $\bar{F}_o = \frac{9(-3a^2c + \sqrt{15}\sqrt{-a^4c^2 + 36a^5c^3})}{8(-2+45ac)}$ otherwise, we show that case (mo) is again the best strategy for the firm. But what is different from Region 2 above is that the outlet store is opened with a distance to the main store, where $t = \hat{t} = \frac{-9a^3c + 2aF_o + 10F_o^2 - 3\sqrt{9a^6c^2 - 4a^4cF_o + a^2F_o^2 - 20a^3cF_o^2}}{10F_o^2}$. Accordingly, the optimal value of other decisions are: for the main store, $(\delta_m = \frac{B}{5c}, p_m = \frac{-20A+7B}{50c})$, which leads to the sales of the main store being $Q_m = 1 - \frac{p_m - p_o - at}{\delta_m - \delta_o} = \frac{40A+B}{5B}$; for the outlet store, $(\delta_o = \frac{B}{10c}, p_o = \frac{3(-10A+B)}{50c})$, which leads to the sales of the outlet store being $Q_o = \frac{p_m - p_o - at}{\delta_m - \delta_o} - \frac{p_o + at}{\delta_o} = \frac{-60A+B}{5B}$. The firm's optimal profit is $\Pi^{mo} = \frac{200A^2 - 10A + B}{25Bc} - (1-t)^2 F_o$, where $A = 20act$ and $B = 1 + \sqrt{1 - 20A}$, and the total profit is composed by the firm's profit in the main store $\Pi_m^{mo} = (p_m - \delta_m^2 c)Q_m = \frac{(5-B)(40A+B)}{250c}$ and its profit in the outlet store profit $\Pi_o^{mo} = (p_o - \delta_o^2 c)Q_o - (1-t)^2 F_o = \frac{(-60A+B)(-10A+B)}{125Bc} - (1-\hat{t})^2 F_o$.

- Otherwise (Region R.1), we show that case (m) dominates the other two cases which implies that the firm will only offer the main store. Accordingly, the optimal decisions are $(\delta_m = \frac{1}{3c}, p_m = \frac{2}{9c})$ which leads to the sales being $Q_m = 1 - \frac{p_m}{\delta_m} = \frac{1}{3}$. The firm's optimal profit is $\Pi^{mo} = \frac{1}{27c}$.

Plotting the three regions derived above on the (F_o, a) panel yields Figure 2 in Proposition 1, which completes the proof of this proposition. By summarizing the optimal decisions and profit in each specific region we derive Table 2 in Proposition 2, which completes the proof of this result.

Proof of Proposition 3: Note that if the outlet store is not offered, the optimal quality and pricing decisions are consistent with those in region R_1. So, we have ($\delta_m = \frac{1}{3c}, p_m = \frac{2}{9c}, Q_m = \frac{1}{3}, \Pi^m = \frac{1}{27c}$). In order to prove this result, we only need to compare the quality and pricing decisions in regions R_2 and R_3 to those values.

- Quality of the main store. From Table 2 of Proposition 2, it is known that the optimal quality in regions R_2 and R_3 is $\delta_m = \frac{1+\sqrt{1-20act}}{5c}$, where $t = 0$ in region R_2 and $t = \hat{t}$ in region R_3. We can verify that $\hat{t} \leq \frac{1}{36ac}$. Given this, it is not difficult to show that

$$\delta_m = \frac{1 + \sqrt{1 - 20act}}{5c} > \frac{1}{3c}.$$

- Price of the main store. In regions R_2 and R_3, the optimal retail price of the main store can be expressed as a function of the quality of the product in the same store, i.e., $p_m = \frac{\delta_m + c\delta_m^2}{2}$. Given $\delta_m > \frac{1}{3c}$ from above, we can easily show that

$$p_m = \frac{\delta_m + c\delta_m^2}{2} > \frac{\frac{1}{3c} + \frac{1}{9c}}{2} = \frac{2}{9c}.$$

- Sales of the main store. The sales volume in the main store is

$$Q_m = \frac{40act + 1 + \sqrt{1 - 20act}}{5(1 + \sqrt{1 - 20act})} = \frac{3}{5} - \frac{2\sqrt{1 - 20act}}{5} < \frac{1}{3},$$

  where the last inequality is due to the fact that $t \leq \hat{t} < \frac{1}{36ac}$.

- Profit in the main store. The profit in the main store is

$$\Pi^{mo} = \frac{(4 - \sqrt{1 - 20act})(1 + 40act + \sqrt{1 - 20act})}{250c},$$

  which is an increasing function in t. So, $\Pi^{mo} < \Pi^{mo}(t = \frac{1}{36ac}) = \frac{1}{27c}$.

Proof of Corollary 2: Since we use the result in Corollary 2 to prove Corollary 1, so we will first prove Corollary 2, followed by the proof of Corollary 1.

Note that this result only applies to the case when both the main and outlet stores are offered so that the distance of the outlet store, measured by t, is relevant. In this case, for any given model parameter set (a, c) and any t, we shall first solve for the optimal quality and pricing decisions for the two stores and express the difference in quality and price in terms of (a, c, t). Accordingly, they can be expressed as follows:

$$\Delta\delta = \delta_m - \delta_o = \frac{1 + \sqrt{1 - 20act}}{10c}; \quad (2.4)$$

$$\Delta p = p_m - p_o = \frac{\delta_m + c\delta_m^2}{2} - \frac{\delta_o + c\delta_o^2 - at}{2} = \frac{2 + 5act + 2\sqrt{1 - 20act}}{25c}. \quad (2.5)$$

Clearly, $\Delta\delta$ in (2.4) is decreasing with t, which implies that quality differentiation decreases in location differentiation, which is measured by t.

Taking derivative of Δp given in (A-2) with respect to t results with

$$\frac{\partial \Delta p}{\partial t} = \frac{a}{5}(1 - \frac{4}{\sqrt{1 - 20act}}) \leq 0,$$

which implies that price differentiation also decreases in location differentiation. This completes the proof of Corollary 2.

Proof of Corollary 1: Recall that this result only applies to region R_3 where the outlet store is active and it is located within a certain distance to the main store. The equilibrium values of qualities, prices and location of the outlet store are given in the last column of Table 2 in the paper.

In terms of the sensitivity analysis on the optimal location of the outlet store $t = \hat{t}$, we have

$$\frac{\partial \hat{t}}{\partial F_o} = \frac{a(18a^2c - 2F_o)G(F_o) + 3a(F_o^2 + 18a^4c^2 - 6a^2cF_o - 20acF_o^2)}{10aF_o^3G(F_o)} \text{ and}$$

$$\frac{\partial \hat{t}}{\partial c} = \frac{3a^2(-3a + \frac{-9a^3c+2aF_o+10F_o^2}{\sqrt{G(F_o)}})}{10F_o^2},$$

where $G(F_o) = \sqrt{9a^4c^2 - 4a^2cF_o + F_o^2 - 20acF_o^2}$. With some simple algebra, we can show that $\frac{\partial \hat{t}}{\partial F_o} \geq 0$ and $\frac{\partial \hat{t}}{\partial c} \geq 0$, which implies that the location differentiation increases in F_o and c.

In terms of sensitivity analysis on the quality differentiation $\delta_m - \delta_o$, we have

$$\frac{\partial(\delta_m - \delta_o)}{\partial F_o} = \frac{\partial(\delta_m - \delta_o)}{\partial \hat{t}} \frac{\partial \hat{t}}{\partial F_o} \leq 0 \text{ and}$$

$$\frac{\partial(\delta_m - \delta_o)}{\partial c} = \frac{2a^2c - F_o + 10acF_o - \sqrt{9a^4c^2 - 4a^2cF_o + F_o^2 - 20acF_o^2}}{10c^2\sqrt{9a^4c^2 - 4a^2cF_o + F_o^2 - 20acF_o^2}}.$$

With some simple algebra, we can show that $\frac{\partial(\delta_m - \delta_o)}{\partial c} \leq 0$ for any (a, c, F_o) in region R_3. This implies that the location differentiation decreases in both F_o and c.

In terms of sensitivity analysis on the price differentiation $p_m - p_o$, we have

$$\frac{\partial(p_m - p_o)}{\partial F_o} = \frac{\partial(p_m - p_o)}{\partial \hat{t}} \frac{\partial \hat{t}}{\partial F_o} \leq 0 \text{ and}$$

$$\frac{\partial(p_m - p_o)}{\partial c} = \frac{\partial(p_m - p_o)}{\partial c}\frac{\partial c}{\partial c} + \frac{\partial(p_m - p_o)}{\partial \hat{t}} \frac{\partial \hat{t}}{\partial c}.$$

With some simple algebra, we can show that $\frac{\partial(p_m - p_o)}{\partial c} \leq 0$, which implies that the price differentiation decreases in both F_o and c.

Proof of Proposition 4: In this proof, we prove items (a) and (b) and we start with (a) first.

(a) In the case without location differentiation, the firm can choose either to open the main store only or to open both the main and outlet stores on the same site.

If the firm chooses to operate the main store only, according to our previous analysis in the proof of Propositions 1 and 2, we have the optimal decisions and profit consistent with the corresponding values in region R_1 of Table 2 of the paper as follows: $\delta_m = \frac{1}{3c}$, $p_m = \frac{2}{9c}$ and $\Pi^m = \frac{1}{27c}$.

If the firm decides to operate both stores, we have the optimal decisions and profit consistent with the corresponding values in region R_2 of Table 2 of the paper as follows: $\delta_m = \frac{2}{5c}$, $\delta_o = \frac{1}{5c}$, $p_m = \frac{7}{25c}$, $p_o = \frac{3}{25c}$ and $\Pi^{mo} = \frac{1}{25c} - F_o$.

By comparing the profits in the above two scenarios, we have $\Pi^{mo} \geq \Pi^m$ if $F_o \leq \frac{2}{675c}$, which implies that the optimal strategy is to open both stores at the same location; otherwise, $\Pi^{mo} \leq \Pi^m$, which implies that the optimal strategy is to open the main store only.

(b) In the case without price and quality differentiation (i.e., there is only location differentiation), we have $\delta_m = \delta_o = \delta$ and $p_m = p_o = p$. Given location differentiation $t > 0$, if consumers buy from the main store, their net utility is $U_m = \delta v - p$. If they buy from the outlet store, the net utility is $U_o = \delta v - p - at$, which is dominated by the net utility of buying from the main store. Therefore, the consumers would never buy from the outlet store and the firm would choose to open the main store only. Accordingly, the optimal decisions are $\delta_m = \frac{1}{3c}$ and $p_m = \frac{2}{9c}$ and the optimal profit is $\Pi^m = \frac{1}{27c}$.

Appendix A (Part II): Base Model with Homogeneous Travel Sensitivity and with Competition from E-commerce

Recall from the main paper that notation γ_m and γ_o represent the switching probability of customers who buy from the main and the outlet stores , respectively, to online channels. Due to competition from e-commerce, the profit function can be revised as follows:

$$\begin{aligned}
\Pi &= (p_m - c_m)Q'_m + (p_o - c_o)Q'_o - (1-t)^2 F'_o - F'_m \\
&= (p_m - c_m)(1-\gamma_m)Q_m + (p_o - c_o)(1-\gamma_o)Q_o - (1-t)^2(1-\gamma_o)F_o - (1-\gamma_m)F_m \\
&= (1-\gamma_m)[(p_m - c_m)Q_m + b(p_o - c_o)Q_o - b(1-t)^2 F_o - F_m],
\end{aligned} \tag{2.6}$$

where $b = \frac{1-\gamma_o}{1-\gamma_m}$ measures the degree of competition from e-commerce. In particular, the higher the value of b is, the higher competition intensity from the e-commerce. Following the profit function, it is clear that the firm's optimal decisions would only depend on b, rather than individual values of switching probabilities, γ_m and γ_o. We follow the same approach as that in the base model without competition from e-commerce. That is, we will analyze the three cases – (m), (o) and (mo) and then compare the firm's profit across these cases to derive the best store offering strategy and the corresponding decisions and profit.

Case (m): In this case, only the main store is available. The firm's profit function can be rewritten as $\Pi = (p_m - c_m)(1-\gamma_m)(1-\frac{p_m}{\delta_m}) = (1-\gamma_m)\left[(p_m - c_m)(1-\frac{p_m}{\delta_m})\right]$, where $\left[(p_m - c_m)(1-\frac{p_m}{\delta_m})\right]$ is identical to the firm's profit function in case (m) without e-commerce. Therefore, the optimal decisions are the same as in case (m) of Lemma 1: $\delta_m = \frac{1}{3c}$, $p_m = \frac{2}{9c}$ for the main store. As for the outlet store, the decision set (t, δ_o, p_o) can be any values that satisfy $p_o \geq \frac{p_m \delta_o}{\delta_m} - at$. Accordingly, the firm's optimal profit is $\Pi = (1-\gamma_m)\frac{1}{27c}$.

Case (o): In this case, only the outlet store is available. The firm's profit function can be rewritten as $\Pi = (p_o - c_o)(1-\gamma_o)(1-\frac{p_o+at}{\delta_o}) - (1-t)^2(1-\gamma_o)F_o = (1-\gamma_o)\left[(p_o - c_o)(1-\frac{p_o+at}{\delta_o}) - (1-t)^2 F_o\right]$, where $\left[(p_o - c_o)(1-\frac{p_o+at}{\delta_o}) - (1-t)^2 F_o\right]$ is identical to the firm's profit function in case (o) without e-commerce. Thus, the optimal decisions are the same as in case (o) of Lemma 1. That is, for the effective outlet store, we have:

- If $F_o \leq \frac{a}{6}$: $(t=0, \delta_o = \frac{1}{3c}, p_o = \frac{2}{9c})$ and $\Pi^{mo} = \frac{1}{27c} - F_o$.

- Otherwise:
  $(t = \frac{a^3c - 2aF_o + 6F_o^2 + \sqrt{a^6c^2 - 4a^4cF_o + a^2F_o^2 + 12a^3cF_o^2}}{6F_o^2}, \delta_o = \frac{1+\sqrt{1+12act}}{6c}, p_o = \frac{1-3act+\sqrt{1+12act}}{9c})$ and $\Pi = \frac{(1-12act+\sqrt{1+12act})^2}{54c(1+\sqrt{1+12act})} - (1-t)^2 F_o$.

As for the inactive main store, the decision set (δ_m, p_m) needs to satisfy $p_o \leq p_m - at + \delta_o - \delta_m$.

Case (mo): As seen from the profit functions in cases with and without e-commerce, in case (mo), the firm's profit functions do not have a proportional relation as that in cases (m) and (o). Due to the additional parameter b in the profit function, we were not able to analytically derive the optimal decisions and profit in this case. So, we use a numerical approach to derive the constrained optimal values in case (mo), followed by a comparison of cases (m), (o) and (mo). Figure 3(a), 3(b) and 3(c) of the main paper presents the impact of e-commerce on i) the firm's optimal store offering strategy, ii) the optimal distance between the two stores, t, and iii) the contribution of the main store to the overall sales, measured by $r = \frac{Q'_m}{Q'_m + Q'_o}$, respectively. In the numerical study, we set $b \in [1, 1.25]$ to focus on the effect of e-commerce and fix $c = 0.1$. In Figures 3(b) and 3(c), we further fix $F_o = 0.05$ and $a = 0.2$, under which both the main and outlet stores are operational with positive sales.

Appendix B: Extended Model with Heterogeneous Travel Sensitivity

Demand function

Let us first derive the demand function. Note that one major difference of this model from the base model with homogeneous sensitivity analysis is that we have two separate customer groups/segments: one with low valuation where $v \in [0, 0.5]$ and the other with high valuation where $v \in [0.5, 1]$. Within each customer group/segment, we could use a similar approach as that used in the base model in §3 to derive the demand function and then we combine the demand functions in the two segments to derive the overall demand.

In segment 1 where customers have low valuations, we could have four possible cases:

- Case (1m): Positive sales occurs in the main store only, which happens when $p_o \geq \frac{\delta_o p_m}{\delta_m} - a_1 t$ and $p_m \leq \frac{\delta_m}{2}$;
- Case (1mo): Positive sales occurs in both the main and outlet store, which happens when $p_m - a_1 t + \frac{\delta_o}{2} - \frac{\delta_m}{2} \leq p_o \leq \frac{\delta_o p_m}{\delta_m} - a_1 t$. These conditions also imply that $p_m \leq \frac{\delta_m}{2}$ and $p_o \leq \frac{\delta_o}{2} - a_1 t$;
- Case (1o): Positive sales occurs in the outlet store only, which happens when $p_o \leq p_m - a_1 t + \frac{\delta_o}{2} - \frac{\delta_m}{2}$ and $p_o \leq \frac{\delta_o}{2} - a_1 t$;
- Case (1n): No sales occurs in either of the two stores, which happens when $p_o \geq \frac{\delta_o}{2} - a_1 t$ and $p_m \geq \frac{\delta_m}{2}$.

Accordingly, the demand function of the main store from segment 1 customers can be summarized as follows:

$$Q_1^m = \begin{cases} \frac{1}{2} - \frac{p_m}{\delta_m} & \text{if in } (1m) \\ \frac{1}{2} - \frac{p_m - p_o - a_1 t}{\delta_m - \delta_o} & \text{if in } (1mo) \\ 0 & \text{if in } (1o) \text{ or } (1n) \end{cases}$$

Similarly, the demand function of the outlet store from segment 1 customers can be summarized as follows:

$$Q_1^o = \begin{cases} 0 & \text{if in } (1m) \text{ or } (1n) \\ \frac{p_m - p_o - a_1 t}{\delta_m - \delta_o} - \frac{p_o + a_1 t}{\delta_o} & \text{if in } (1mo) \\ \frac{1}{2} - \frac{p_o + a_1 t}{\delta_o} & \text{if in } (1o) \end{cases}$$

As for segment 2 customers, we again have four possible cases:

- Case (2m): Positive sales occurs in the main store only, which happens when $p_o \geq \frac{\delta_o p_m}{\delta_m} - a_2 t$;
- Case (2mo): Positive sales occurs in both the main store and the outlet store, which happens when $p_m - a_2 t + \delta_o - \delta_m \leq p_o \leq \frac{\delta_o p_m}{\delta_m} - a_2 t$ and $p_m - a_2 t + \frac{\delta_o}{2} - \frac{\delta_m}{2} \geq p_o$;
- Case (2o): Positive sales occurs in the outlet store only, which happens when $p_o \leq p_m - a_2 t + \delta_o - \delta_m$;
- Case (2n): No sales occurs in either of the two stores, which happens when $p_o \geq \delta_o - a_2 t$ and $p_m \geq \delta_m$.

Accordingly, the demand function of the main store from segment 2 customers can be summarized as follows:

$$Q_2^m = \begin{cases} 1 - max\{\frac{p_m}{\delta_m}, \frac{1}{2}\} & \text{if in } (2m) \\ 1 - max\{\frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}, \frac{1}{2}\} & \text{if in } (2mo) \\ 0 & \text{if in } (2o) \text{ or } (2n) \end{cases}$$

Similarly, the demand function of the outlet store from segment 2 customers can be summarized as follows:

$$Q_2^o = \begin{cases} 0 & \text{if in } (2m) \text{ or } (2n) \\ max\{\frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}, \frac{1}{2}\} - max\{\frac{p_o + a_2 t}{\delta_o}, \frac{1}{2}\} & \text{if in } (2mo) \\ 1 - max\{\frac{p_o + a_2 t}{\delta_o}, \frac{1}{2}\} & \text{if in } (2o) \end{cases}$$

Since there are four possible cases for each segment, we totally have 16 possible cases. However, some of the cases are not possible. For example, due to the assumption $p_m \leq \delta_m$, case (2n) is not possible. Also segment 2 customers have a higher valuation than segment 1 customers, which implies that whenever segment 1 customers buy from the main store, segment 2 customers would also buy from the main store. In other words, when case (1m) happens, case (2m) would also happen. Therefore, we could rule out the combination of cases (1m) and (2o) and the combination of cases (1mo) and (2o). As a result, there are 10 possible combinations of cases in total. They are:

1) Combination of case (1n) and case (2m);

2) Combination of case (1m) and case (2m): This combination requires $\frac{p_m}{\delta_m} \leq \frac{1}{2}$. Accordingly, the profit function can be written as: $\Pi = (p_m - c\delta_m^2)(\frac{1}{2} + \frac{1}{2} - \frac{p_m}{\delta_m}) - F_o(1-t)^2$. It is straightforward to characterize the optimal decisions as: $p_m = \frac{2}{9c}$ and $\delta_m = \frac{1}{3c}$. As a result, we have $\frac{p_m}{\delta_m} = \frac{2}{3} \geq \frac{1}{2}$, which contradicts with the condition of $\frac{p_m}{\delta_m} \leq \frac{1}{2}$. So, this case is not optimal.

3) Combination of case (1mo) and case (2m): This combination requires $\frac{p_m}{\delta_m} \leq \frac{1}{2}$. Accordingly, the profit function can be written as: $\Pi = (p_m - c\delta_m^2)(\frac{1}{2} + \frac{1}{2} - \frac{p_m - p_o - a_1 t}{\delta_m - \delta_o}) + (p_o - c\delta_o^2)(\frac{p_m - p_o - a_1 t}{\delta_m - \delta_o} - \frac{p_o + a_1 t}{\delta_o}) - (1-t)^2 F_o$. It is straightforward to derive the optimal decisions as: $p_m = \frac{1}{2}(\delta_m + c\delta_m^2)$ and $\delta_m = \frac{1+\sqrt{1-20a_1 ct}}{5c}$.

Since $\frac{p_m}{\delta_m} = \frac{1}{2}(1 + c\delta_m) \geq \frac{1}{2}$, which contradicts with the condition $\frac{p_m}{\delta_m} \leq \frac{1}{2}$. So, this case is not optimal either.

4) Combination of case (1o) and case (2m);

5) Combination of case (1n) and case (2mo);

6) Combination of case (1m) and case (2mo): This combination requires $\frac{\delta_o p_m}{\delta_m} - a_1 t \leq p_o$ and $p_m - a_2 t + \delta_o - \delta_m \leq p_o \leq \frac{\delta_o p_m}{\delta_m} - a_2 t$. These conditions further lead to $\frac{p_o + a_2 t}{\delta_o} \leq \frac{p_m}{\delta_m} \leq \frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}$. Since $a_1 \leq a_2$, we also have $\frac{p_o + a_1 t}{\delta_o} \leq \frac{p_o + a_2 t}{\delta_o} \leq \frac{p_m}{\delta_m}$, which leads to $\frac{\delta_o p_m}{\delta_m} - a_1 t \geq p_o$, contradictory to the requirement of this combination. So, this case is not possible either.

7) Combination of case (1mo) and case (2mo): This combination requires $\frac{p_m - p_o - a_1 t}{\delta_m - \delta_o} \leq \frac{1}{2} \leq \frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}$. However, due to $a_1 \leq a_2$, we must have $\frac{p_m - p_o - a_1 t}{\delta_m - \delta_o} \geq \frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}$. Contradiction. Therefore, this case is not possible either.

8) Combination of case (1o) and case (2mo);

9) Combination of case (1n) and case (2o): In Proposition 1 of the main paper, we have proved that case (2o) is dominated by case (2mo). So, this case is not optimal either.

10) Combination of case (1o) and case (2o): This combination is dominated by combination (8), since it can be viewed as a special case of combination (8) when one of the constraints is binding, resulting in $Q_2^m = 0$. So, this case is not optimal either.

According to the analysis above, there are only four combinations/cases that can possibly be optimal: (1) combination of case (1n) and case (2m); (2) combination of case (1o) and case (2m); (3) combination of case (1n) and case (2mo); and finally, (4) combination of case (1o) and case (2mo). For presentation purpose, possible scenarios of the demand function for the main store have been summarized in Table 4 in the main paper. Further, the demand functions for the four combinations that can possibly be optimal have been summarized in Table 5 in the main paper. The four combinations here correspond to the four cases presented in Table 5.

The next step is to analyze each of the four possible combinations/cases and compare to derive the best combination.

Case (1): In this combination, only segment 2 consumers would buy from the main store. This is corresponding to case (m) in the base model studied in §4. So, the optimal decisions are the same as that of case (m) in Lemma 1, where $\delta_m = \frac{1}{3c}$ and $p_m = \frac{2}{9c}$. Accordingly, the firm's optimal profit is $\Pi = \frac{1}{27c}$.

Case (2): In this combination, segment 1 customers would buy from the outlet store and segment 2 customers would buy from the main store. The profit function is $\Pi = (p_m - c\delta_m^2)(1 - \frac{p_m}{\delta_m}) + (p_o - c\delta_o^2)(\frac{1}{2} - \frac{p_o + a_1 t}{\delta_o}) - (1-t)^2 F_o$. Solving the first-order conditions yields a stationary point that can possibly be optimal:

$$\begin{aligned}
\delta_m &= \frac{1}{3c}, \quad p_m = \frac{2}{9c} \\
t &= \frac{2a_1^3 c - 2a_1 F_o + 12F_o^2 + a_1\sqrt{4a_1^4 c^2 - 8a_1^2 c F_o + F_o^2 + 48 a_1 c F_o^2}}{12F_o^2} \\
\delta_o &= \frac{1 + \sqrt{1 + 48 a_1 c t}}{12c}, \quad p_o = \frac{1 - 12 a_1 c t + \sqrt{1 + 48 a_1 c t}}{36c}.
\end{aligned} \tag{2.7}$$

Case (3): In this combination, segment 2 customers would buy from both the main and the outlet stores. This is corresponding to case (mo) in the base model studied in §4. The optimal unconstrained decisions are the same as that of case (mo) in Lemma 1:

$$t = \frac{-9a^3c + 2aF_o + 10F_o^2 - 3\sqrt{9a^6c^2 - 4a^4cF_o + a^2F_o^2 - 20a^3cF_o^2}}{10F_o^2}$$
$$\delta_m = \frac{B}{5c}, p_m = \frac{-20A + 7B}{50c}, \delta_o = \frac{B}{10c}, p_o = \frac{3(-10A + B)}{50c},$$

where $A = 20act$ and $B = 1 + \sqrt{1 - 20A}$.

Case (4): In this combination, segment 1 customers would buy from the outlet store, segment 2 customers would buy from both the main store and the outlet stores. The profit function is $\Pi = (p_m - c\delta_m^2)(1 - \frac{p_m - p_o - a_2 t}{\delta_m - \delta_o}) + (p_o - c\delta_o^2)(\frac{p_m - p_o - a_2 t}{\delta_m - \delta_o} - \frac{p_o + a_2 t}{\delta_o} + \frac{1}{2} - \frac{p_o + a_1 t}{\delta_o}) - (1-t)^2 F_o$. Unfortunately, we were not able to solve the first-order conditions for all five decision variables. However, if we treat δ_o as fixed, solving the other four first-order conditions yields a stationary point that can possibly be optimal:

$$\delta_m = \frac{C - \sqrt{C^2 + 3c\delta_o(a_1a_2 - a_2^2 - \delta_oF_o)\left[2a_1^2c\delta_o - 2(4a_2 + \delta_o + 2c\delta_o^2)F_o + a_1(a_2 - 2a_2c\delta_o + 4F_o)\right]}}{6c(a_1a_2 - a_2^2 - \delta_oF_o)} \quad (2.8)$$
$$t = \frac{\delta_o(-2a_1\delta_m + 2a_2\delta_m - 2a_1c\delta_m^2 + 4a_2c\delta_m^2 - a_1\delta_o + 4a_1c\delta_m\delta_o - 4a_2c\delta_m\delta_o + 2a_1c\delta_o^2 + 8\delta_mF_o + 8\delta_oF_o)}{2(-4a_1a_2\delta_m + 4a_2^2\delta_m - a_1^2\delta_o + 4\delta_m\delta_oF_o + 4\delta_o^2F_o)}$$
$$p_m = \frac{\delta_m(2\delta_m + 2c\delta_m^2 + \delta_o + 2c\delta_o^2 - 2a_1t + 4a_2t)}{4(\delta_m + \delta_o)},$$
$$p_o = \frac{2\delta_m\delta_o + 2c\delta_m^2\delta_o + \delta_o^2 + 2c\delta_o^3 - 4a_2\delta_mt - 2a_1\delta_ot}{4(\delta_m + \delta_o)},$$

where $C = a_1a_2 - a_2^2 - a_1^2c\delta_o - \delta_o(1 - 4c\delta_o)F_o$.

Given the complication in the firm's profit function in δ_o, it is analytically challenging to derive the unconstrained optimal δ_o in this case. Hence, we resort to numerical analysis to obtain the constrained optimal δ_o and the other corresponding decisions and profit. Without loss of generality, we have used $c = 0.1$ throughout the numerical study here. We will focus on the other three parameters F_o, a_1 and a_2. In particular, our numerical study indicates that the effect of travel sensitivity heterogeneity depends on whether F_o is below or above a threshold value, $F_o' = \frac{2}{675c}$, which also appears in the base model in Figure 2. Hence, the two values we take for F_o are $(\frac{19}{20})(\frac{2}{675c}) < F_o'$ and $(4)(\frac{2}{675c}) > F_o'$ which represent cases of low and high fixed cost of the outlet store (compared to the threshold F_o'), respectively.

Similar to the case with homogeneous travel sensitivity, we can plot the firm's optimal store strategy (whether to operate both the main and outlet stores) in the (a_1, a_2) plane. In our numerical examples, the travel sensitivity a_1 and a_2 range from 0 to 1, where $a_2 \geq a_1$. So we only need to examine the upper left triangle. Moreover, due to different strategies derived under a low or a high fixed cost for the outlet store, we plot the case with a low fixed cost in Figure 4(a) and the case with a high fixed cost in Figure 4(b) in the main paper.

Chapter 3

Coordinating Supply and Demand on an On-demand Service Platform with Impatient Customers

3.1. Introduction

Recent advances in internet/mobile technologies have enabled the creation of various innovative on-demand service platforms for providing *on-demand* services anytime/anywhere. Examples include grocery delivery services (e.g., Instacart, Google Express), meal delivery services (e.g., Sprig, Blue Apron), and food delivery services directly from restaurants (e.g., DoorDash, Deliveroo (U.K.), Yelps Eat24), consumer goods delivery services (e.g., UberRush), dog-walking services (e.g., Wag), and taxi-style transportation (e.g., Uber, Didi).

To meet dynamic customer demand anytime/anywhere, it is economical for on-demand service firms to use independent providers (or agents) to fulfill customer requests quickly.

However, using independent agents to deliver on-demand services can be challenging because work participation of independent providers is primarily driven by earnings. However, because independent agents do not get compensated for idle times, earnings depends on wage rate and utilization, which depends on customer demand. At the same time, the demand associated with wait-time and price sensitive customers depends on two key factors: price and waiting time. Since customer's waiting time depends on the number of participating agents, which depends on the wage and the customer demand. Therefore, the "supply" of participating agents and the "demand" of customer requests are endogenously dependent on the wage and the price specified by the firm.

The underlying interactions between supply and demand through wage and price selections make it challenging for an on-demand service firm to coordinate endogenous supply and demand in different time periods by: (1) setting the right wage (i.e., compensation) to get the right supply (i.e., the right number of earnings sensitive participating agents); and (2) charging the right price to control the right demand (i.e., the right amount of wait-time and price sensitive customers). To elaborate, consider the simple case when the demand is fixed. If the firm offers a higher wage, more agents will participate and customer satisfaction will increase due to a quicker service. However, participating agents will earn less due to low utilization. On the other hand, if the firm offers a lower wage, fewer agents will participate and customer satisfaction will decrease due to longer waiting times. For example, as reported by Klein (2016), the recent closure of SpoonRocket (a 10-minute meal on-demand delivery service based in Berkeley, California) was due to the low wages offered by SpoonRocket to its independent drivers that resulted in an insufficient number of independent participating drivers. Consequently, many meals were delivered late, causing many unhappy customers and sales dropped subsequently. Eventually, SpoonRocket was bankrupt in March 2016.

In view of the intricate relationship between endogenous supply and demand through wage and price selections, we develop an analytical framework to examine how an on-demand service firm should set its price, wage and payout ratio (i.e., the ratio of wage over price). (Throughout this paper, we shall refer to "payout ratio" as the percentage of the price collected from the customers that is paid to the providers.) In our framework, we use a queueing model to study the situation where both supply (i.e., number of providers) and demand (i.e., customer arrival rate) are "endogenously" dependent on wage, price and other operating factors. Our model captures an operating environment where (1) wait-time and price sensitive customers are "heterogeneous" in their *evaluation* of the service; and (2) earnings sensitive independent providers are "heterogeneous" in their *reservation price* (i.e., the minimum wage for work participation).

By analyzing the steady state performance of our queueing model in equilibrium, we characterize the optimal price, wage and payout ratio (i.e., the ratio of wage over price) in the basic setting under which the objective is to maximize the firm's profit. We then extend our analysis to a more general setting under which the objective is to maximize the firm's profit plus the social (customer and provider) welfare. For both settings, we obtain two key findings:

1. When the potential customer demand becomes higher, it is optimal for the firm to charge a higher price, pay a higher wage, and offer a higher payout ratio.
2. When customers become more wait-time sensitive, it is optimal for the firm to pay a higher wage and offer a higher payout ratio; however, the firm may need to charge a lower price to sustain the demand of increasingly impatient customers.

Our findings have the following managerial implications. First, as both the optimal price and the optimal wage are increasing in the maximum potential customer demand rate, our result provides an additional explanation/justification for an on-demand service firm (such as Uber) to charge its customers a higher price and pay its independent providers a higher wage when demand is higher. Second, while it is simple to share a fixed percentage of its revenue with the independent agents (e.g., Uber shares 80% of its revenue with its drivers; see Damodaran (2014)), we find that the firm can increase its own profit as well as the total (customer and provider) welfare by offering a higher payout ratio when demand is higher. We hope this result might motivate on-demand service firms to re-evaluate their current fixed revenue sharing scheme. For instance, the firm may offer a higher (lower) payout ratio during peak hours (non-peak hours). Third, we also find that it is optimal for the firm to reduce its payout ratio when the number of registered independent providers becomes larger. This analytical result provides an economic justification for explaining why Uber reduced its payout ratio from 0.8 (initial payout ratio for its first cohorts of drivers) to 0.75 (for its second cohorts of drivers in 2014). Fourth, for urgent on-demand services with highly wait-time sensitive customers, the firm may need to lower its price to sustain demand from increasingly impatient customers.

This paper is organized as follows. We provide a brief review of related literature in Section 2. Section 3 presents our queueing model of endogenous supply and demand along with heterogeneous providers and customers. In Section 4, we analyze the equilibrium behavior of our queueing system to determine the optimal price, wage and payout ratio for maximizing the firm's profit. We also adapt our base model to two special cases when the firm uses a fixed payout ratio and when the firm sets a fixed service level. In Section 5, we extend our analysis to the case when the objective of the firm is to maximize its own profit plus the total (customer and provider) welfare. We construct some illustrative numerical examples in Section 6 based on actual data provided by Didi: the leading taxi-style transportation on-demand service in China. We conclude the paper in Section 7. We also extend some of our analytical results for the base model under more general assumptions in Appendix A. For ease of exposition, all mathematical proofs for the results in the main text are provided in Appendix B.

3.2. Literature Review

Our paper belongs to an emerging stream of research that examines operations and pricing issues arising from the *sharing economy*. Besides on-demand service platforms, a number of researchers have also studied other types of platforms in the sharing economy. Benjaafar et al. (2015), Fraiberger and Sundarajan (2015), and Jiang and Tian (2015) examined a customer's decision to purchase or to rent assets in the presence of "product sharing platforms" such as Airbnb. By crawling data from Airbnb, Li et al. (2015) showed empirically that "professional" owners earned more. Our paper differs from this stream of research in the following aspect. While product sharing platforms set the payout amounts to the owners and owners set the price, customers using such product sharing platforms often reserve the service in advance, which present very different issues in the decision making process and timing of the underlying service request mechanisms, as compared with on-demand service platforms which provide wait-time sensitive service on-demand.

Recent developments of various on-demand service platforms such as Uber and DoorDash (see Kokalitcheva (2015), Wirtz and Tang (2016), and Shoot (2015)) have motivated researchers to explore various operational issues. First, there is an on-going debate regarding the definition of independent contractors for various on-demand service platforms (e.g., see Roose (2014)). At the same time, it is of interest to examine how dynamic wage affects supply especially when independent providers can freely choose whether and when to work. Chen and Sheldon (2015) examined transactional data associated with 25 million trips obtained from Uber and showed empirically that dynamic wage (due to surge pricing) could entice independent drivers to work for longer hours. Moreno and Terwiesch (2014) also examined empirically the independent contractor's bidding behavior on freelancing platforms. Allon et al. (2012) explored the process for matching providers to consumers when capacities are exogenous.

A number of researchers have recently studied the impact of wage and price on supply and demand in the context of on-demand services. Specifically, they examined whether it would be beneficial for an on-demand service firm to adjust its prices and wages dynamically based on real-time system status including the current number of customers requesting service and the number of providers in the system. Riquelme et al. (2015) and Cachon et al. (2015) compared the impact of static versus dynamic prices and wages. By assuming that customers are heterogeneous in terms of valuation and the payout ratio is exogenously given, Riquelme et al. (2015) found that static pricing performs well. On the other hand, Cachon et al. (2015) found that surge pricing performs well by assuming that customers are homogeneous and the payout ratio is endogenously determined. Hu and Zhou (2016) developed a general model where supply purely depends on wage and demand purely depends on price, and derived the conditions under which the optimal revenue sharing ratio is a linear function of the demand rate. Gurvich et al. (2015) also developed a newsvendor-style model to examine the optimal price and wage decisions. This stream of research assumes that customer demand is independent of waiting time and supply (or capacity) is independent of system utilization over time. In contrast, our model captures the rational behavior of customers who are sensitive to wait-time (and price) and independent providers who are sensitive to earnings which depend on the system utilization.

One research stream in the queueing literature has studied pricing decisions for services where customers can incur waiting or delay costs. Of particular relevance to our paper, a number of research papers have examined an operating environment that uses a static uniform (non-discriminatory) pricing strategy for heterogeneous customers. Afeche and Mendelson (2004) analyzed the revenue-maximizing and socially optimal equilibria under uniform pricing for heterogeneous customers with different evaluations of their service, and found that the classical result that the revenue-maximizing admission price is higher than the socially-optimal price (e.g., see Naor (1969)) can be reversed under a more generalized delay cost structure. Zhou et al. (2014) analyzed the structure of the optimal uniform pricing strategies for two classes of customers with different service valuations and wait-time sensitivities. Armony and Haviv (2003) and Afanasyev and Mendelson (2010) studied the competition between two firms under uniform pricing for two classes of heterogeneous customers. All the above research papers, however, are based on the assumption that capacity is exogenously given. In contrast, our paper considers the case when the supply (capacity) depends on wage and system utilization, which needs to be determined endogenously.

Finally, our model is closely related to some recent work by Taylor (2016). To our knowledge, Taylor (2016) is the first to examine pre-committed price and wage based on customer demand and other operating factors in the context of on-demand services. He compared the optimal prices when the providers are independent contractors or regular employees, and examined the impact of wait-time sensitivity on the optimal price and wage using a two-point distribution for both the customer valuation of the service and the provider's reservation price. Our model allows these two distributions to be continuous, and complements Taylor's work in two important ways. First, our focus is to examine the impact of demand rate, wait-time sensitivity, service rate, and the size of available providers (who are on-reserve) on the optimal price, wage and payout ratio (ratio between the optimal wage and the optimal price). Second, in addition to maximizing its profit, we also consider the case when the firm maximizes the sum of its own profit and the total welfare. We find that our key results continue to hold: the optimal price, the optimal wage and the optimal revenue sharing ratio are increasing in the potential customer demand rate.

3.3. A Model of Wait-time Sensitive Demand and Earnings Sensitive Supply

We consider an on-demand service platform that coordinates randomly arriving (price and wait-time sensitive) customers with (earnings sensitive) independent service providers. As our motivating example, we shall use on-demand transportation service platforms (such as Uber) to illustrate our model formulation and results throughout this paper, whereas our model can also be used to study other on-demand service applications.

Customers arrive randomly at the platform to request for service, and each service request consists of a (random) amount of service units to be processed by a service provider (e.g., travel distance in km). Throughput this paper, we assume that the requested service by any customer can be met by any of the available service providers. The platform charges each customer a *fixed price rate* p per service unit (e.g., dollar per km), and offers a *fixed wage rate* w per service unit to each participating service provider. (Here, we use "wage rate" per service unit so that the payout ratio $\frac{w}{p}$ is well defined. However, we shall compute "earnings rate" per unit time later for providers to decide whether to participate or not.)

In the same spirit as in Taylor (2016), the price rate p and wage rate w are pre-committed. However, their values can vary across different time periods depending on the specific market characteristics such as the average customer demand rate and the expected number of available providers. In other words, we focus on time-based price/wage over peak/non-peak periods instead of real-time dynamic pricing that depends on real-time system status including the number of customers requesting service and number of available providers in real time.[1]

[1] As articulated in MacMillan (2015) and Taylor (2016), other than Uber and Lyft, many customers resist real time dynamic pricing and most on-demand service providers tend to adopt this form of time-based pricing.

Each customer decides whether to use the platform to request for service, and each independent provider decides whether to participate. We assume that the price rate p and wage rate w are known to the customers and the providers in advance so that they can make their informed decisions. For each service request, the platform will assign one of the available participating providers to serve the customer.[2] The primary objective of the service platform is to select the optimal price rate and wage rate, denoted by p^* and w^*, so as to maximize its average profit.

3.3.1 Customer request rate λ and price rate p

Consider a certain time period (say, peak hours from 8am to 10am). The maximum potential customer demand rate for the service during this time period is given by $\bar{\lambda}$, each of which has a valuation of the service that is based on a value rate v per service unit, where v varies across customers. To model heterogeneous customers without losing tractability, we assume that there is a continuum of customer types so that the value rate v spreads over the range $[0, 1]$ according to a cumulative distribution function $F(.)$, where $F(.)$ is a strictly increasing function with $F(0) = 0$ and $F(1) = 1$.

To capture the notion of wait-time sensitivity, we assume that the utility function of a customer with value rate v is given by

$$U(v) = (v - p)d - cW_q, \tag{3.1}$$

where:

$(v - p)$ is the surplus per service unit and d represents the average service units dictated by a

[2] Our model does not consider any specific assignment mechanism. For instance, the service platform can assign an available participating provider based on certain specific criteria (e.g, Uber assigns an available driver closest to the pickup location), or can announce a service request to all available participating service providers and assign the request to the first respondent.

customer (not the provider)[3], c denotes the cost of waiting per unit time, and W_q represents the expected wait-time for the service.[4]

Observe from (3.1) that $U(v) = (v - p)d - cW_q$ and assume that a rational customer with valuation v will request for service only if $U(v) \geq 0$,[5] the platform can use p and w to indirectly control the effective demand (i.e., the customer request rate) λ so that

$$\lambda = Prob\{U(v) \geq 0\} \cdot \bar{\lambda} = Prob\{v \geq p + \frac{c}{d}W_q\} \cdot \bar{\lambda}.$$

By defining the "target" service level $s = Prob\{v \geq p + \frac{c}{d}W_q\}$, the effective customer request rate λ (i.e., demand) is given by:

$$\lambda = s\bar{\lambda}. \tag{3.2}$$

Because of the one-to-one correspondence between target service level s and the effective demand rate λ, we shall focus our analysis on s instead of λ throughout this paper for mathematical convenience. Using the fact that $s = Prob\{v \geq p + \frac{c}{d}W_q\}$ and that $v \sim F(.)$, the price rate p satisfies the following equation:

$$p = F^{-1}(1 - s) - \frac{c}{d}W_q, \tag{3.3}$$

where the price rate p decreases in the expected wait-time W_q and the unit waiting cost c.

[3]By leveraging internet and mobile technologies, customer requests (e.g., pick up and drop off locations) and the service operations (e.g., route) can be monitored or controlled by the on-demand platform. As such, service units (e.g., travel distance) can be assumed to be dictated by the customers, and cannot be manipulated by the service providers. Thus, the service providers cannot increase their earnings by lengthening the travel distance deliberately with information transparency and real-time location tracking capabilities.

[4]We assume that customers can easily acquire knowledge about the expected waiting time for service based on their prior experience or through internet/social media. For example, as reported by Mosendz and Sender (2014), the average waiting time for an Uber service in different major cities (New York City, San Francisco, etc.) is between 3 to 4 minutes.

[5]In other words, in equilibrium, only customers with value rate $v \geq p + \frac{c}{d}W_q$ will use the platform to request for service, and customer requests with value rate $v < p + \frac{c}{d}W_q$ will not use the platform to meet their service need.

3.3.2 Number of participating providers k and wage rate w

Let K be the (maximum) number of potential earnings-sensitive providers who may decide to participate over the same time period. (Essentially, K represents the number of registered providers who are eligible to participate.) For any given (p, w), let k be the actual number of providers participating on the platform, where $k \leq K$. Also, let μ denote the average service speed (number of service units processed per unit time; e.g., travel speed measured in terms of km per hour) of the service providers so that μ/d represents the service rate of the providers (i.e., average number of customers served per hour).[6] Given the customer request rate λ and the number of participating providers k, the utilization of these k participating providers is equal to $\frac{\lambda}{k\cdot(\mu/d)}$, where $\lambda d < k\mu$ to ensure system stability. The average wage per unit time of a participating provider (when working) is equal to the wage per service unit w multiplied by the average service speed μ. Accounting for the utilization, the average "earning rate" per unit time of a participating provider is equal to $w\mu \cdot \frac{\lambda d}{k\mu} = w\frac{\lambda d}{k}$.[7]

To model the notion of earnings-sensitivity, we assume that each potential provider has a reservation rate r per unit time (i.e., corresponding to his outside option), where r varies across different providers. To model the heterogeneity among providers, we assume that there is a continuum of provider types so that the reservation rate r spreads over the range $[0, 1]$ according to a cumulative distribution function $G(.)$, where $G(.)$ is a strictly increasing function with $G(0) = 0$ and $G(1) = 1$. For a (potential) provider with reservation rate r, he will participate to offer service only if his average earning rate $w\frac{\lambda d}{k}$ is at least equal to r.

Let β denote the proportion of providers who participate in the platform to offer service during this time period. Then, $\beta = Prob\{r \leq w\frac{\lambda d}{k}\} = G(w\frac{\lambda d}{k})$, and the actual number of

[6]If the service units d are already measured in terms of time units, we can simply set $\mu = 1$ in this case.

[7]For independent service providers, utilization and wage rate are the two key factors for their participation. For example, Depills (2016) reported that Uber drivers obtain higher earnings primarily because their utilization rate (measured in terms of percentage of miles driven with a passenger) is much higher than that for taxi drivers. For instance, Uber driver's utilization is 64.2%, while taxi driver's utilization is only 40.7% in Los Angeles.

participating providers k (i.e., supply) is given by

$$k = \beta K. \tag{3.4}$$

Also, in equilibrium, $\beta = G(w\frac{\lambda d}{k})$ so that:

$$G^{-1}(\beta) = w\frac{\lambda d}{k}. \tag{3.5}$$

From (3.4) and (3.5), we can express the wage rate w as a function of the number of participating providers k:

$$w = G^{-1}(\beta)\frac{k}{\lambda d} = G^{-1}(\frac{k}{K})\frac{k}{\lambda d}. \tag{3.6}$$

3.3.3 Problem Formulation

Since the platform earns an average profit of $(p-w)d$ for each customer request, the platform's average total profit is then equal to $\pi = \lambda(p-w)d$. By substituting (3.3) and (3.6) into the profit function, we can express the profit function π as a function of (k, s) below:

$$\pi(k, s) = \lambda d \left[F^{-1}(1-s) - \frac{c}{d}W_q - G^{-1}(\frac{k}{K})\frac{k}{\lambda d} \right]. \tag{3.7}$$

Considering the system stability condition $\lambda d < k\mu$, the optimization problem of the platform can be formulated as

$$\max_{k,s} \ \pi(k, s) \equiv \lambda d \left[F^{-1}(1-s) - \frac{c}{d}W_q - G^{-1}(\frac{k}{K})\frac{k}{\lambda d} \right], \text{ subject to } k > \frac{\lambda d}{\mu},$$

from which we can determine the optimal supply (i.e., the number of participating providers k^*) and the optimal demand (i.e., λ^* via optimal s^* through (3.2)). Then, we can use (3.3) and (3.6) to retrieve the corresponding optimal price rate p^* and optimal wage rate w^* from k^* and λ^*.

3.3.4 Notation

This summarizes our model and problem formulation that captures the impact of price and wage rates on the decisions of wait-time and price sensitive customers to request for service and earnings sensitive providers to participate. For ease of reference, we list below the basic notation used in the paper.

- K : Maximum number of potential service providers who may opt to participate;
- k : Actual number of participating service providers ($k \leq K$);
- $\bar{\lambda}$: Customer demand rate who may opt to use the platform to request for service;
- λ : Actual customer request rate ($\lambda \leq \bar{\lambda}$);
- d : Average amount of service units per service request;
- μ : Average service speed of the service providers;
- v : Value rate per service unit of a customer;
- $F(.)$: Cumulative distribution of value rate of customers v;
- r : Reservation (earning) rate of service providers;
- $G(.)$: Cumulative distribution of reservation rate of service providers r;
- c : Unit waiting cost of customers;
- s : Target service level;
- p : Price rate (price per service unit) charged to customers;
- w : Wage rate (wage per service unit) paid to service providers.

3.4. The Base Model

To characterize the optimal price and optimal wage, we need to determine the joint optimal values of (s, k) that maximize the expected profit $\pi(k, s)$ given in (3.7) subject to the system stability constraint: $k > \frac{\lambda d}{\mu}$. To explicate our analysis, we shall assume that the distribution of value rate v and the reservation wage rate r are uniformly distributed over the range [0,1] so that $F(v) = v$ and $G(r) = r$. Furthermore, we shall approximate the (expected) waiting time W_q given in the customer's utility function (3.1) based on an $M/M/1$ queue with service rate $k(\frac{\mu}{d})$ so that the wait-time function W_q has the following simple closed-form expression:

$$W_q = \frac{\lambda}{\left(k \cdot \frac{\mu}{d}\right) \cdot \left(k \cdot \frac{\mu}{d} - \lambda\right)} = \frac{\lambda d^2}{k\mu(k\mu - \lambda d)}. \tag{3.8}$$

While we shall assume that the following assumption holds for the reminder of this paper, all our results can be directly extended to the more general case where the positive support of the uniform distribution of $F(.)$ or $G(.)$ is within the ranges of $[a, b]$ rather than $[0, 1]$, as used in our illustrative numerical examples in Section 6.

Assumption 1: *$F(.) \sim U[0, 1]$, $G(.) \sim U[0, 1]$, and $W_q = \frac{\lambda d^2}{k\mu(k\mu - \lambda d)}$.*

Under Assumption 1, the price, wage and profit functions given in (3.3), (3.6) and (3.7), respectively, can be simplified as:

$$p = (1 - s) - c\left(\frac{\lambda d}{k\mu}\right)\frac{1}{k\mu - \lambda d} \tag{3.9}$$

$$w = \frac{k^2}{K\lambda d} \tag{3.10}$$

$$\pi(k, s) = \lambda d\left[(1 - s) - c\left(\frac{\lambda d}{k\mu}\right)\frac{1}{k\mu - \lambda d} - \frac{k^2}{K\lambda d}\right]. \tag{3.11}$$

By using the above expressions, we can maximize the expected profit $\pi(k, s)$ given in (3.7) subject to the system stability constraint: $k > \frac{\lambda d}{\mu}$, and obtain the following results:

Proposition 1. *The optimal price p^*, the optimal wage w^*, and the platform's optimal profit π^* exhibit the following characteristics:*

1. *When K or μ increases, w^* decreases, π^* increases, but p^* is not necessarily monotonic.*

2. *When c increases, w^* increases, π^* decreases, but p^* is not necessarily monotonic.*

3. *When $\bar{\lambda}$ or d increases, w^*, p^* and π^* increase.*

As given in the proof of Proposition 1, we can also derive some monotonicity properties on how the different model parameters affect the optimal service level s^*, the optimal number of providers k^*, the optimal expected wait-time W_q^*, the optimal customer request rate λ^*, and the optimal system utilization $\rho^* = \frac{\lambda^* d}{k^* \mu}$. We summarize these monotonicity properties in Table 1.

Table 3.1: Impact of model parameters on s^*, k^*, W_q^*, λ^* and ρ^*.

	s^*	k^*	W_q^*	λ^*	ρ^*
K	↑	↑	↓	↑	×
μ	↑	×	↓	↑	×
c	↓	×	↓	↓	↓
λ	↓	↑	↑	↑	↑
d	↓	↑	↑	↓	↑

↑(increasing); ↓(decreasing); ×(non-monotonic)

Proposition 1 shows that when the maximum number of potential providers K (or when the service speed μ) increases, the potential capacity of the system becomes larger. As such, the platform can increase the number of providers k^* and increase the service rate s^* (or the corresponding customer request rate λ^*) by lowering its wage rate w^*, and can obtain a higher profit π^*. However, when k^* and s^* (as well as λ^*) increase, Equation (3.9) reveals that the optimal price rate p^* is not necessarily monotonic. This explains the first statement. This result implies that it is beneficial for the platform recruit more potential service providers K to join the platform, and help (if possible) to increase their average service speed μ.

Next, when the waiting cost c increases, the platform should lower its the service level s^* so as to reduce the corresponding customer request rate λ^* and expected wait-time W_q as given in (3.8). Consequently, the platform earns less. However, as the optimal number of providers k^* is not necessarily monotonic, Equation (3.9) reveals that the optimal price rate p^* is also not necessarily monotonic. This explains the second statement.

Finally, when the potential customer demand rate $\bar{\lambda}$ increases, the third statement reveals that the platform should increase its price rate p^* to increase the customer request rate λ^* (even though the service level s^* is actually lower since $\lambda^* = s^* \cdot \bar{\lambda}$), and increase its wage rate w^* so as to attract more providers k^* to participate. Overall, the platform earns a higher profit π^* when the potential customer demand rate $\bar{\lambda}$ increases. Also, when the average amount of service units d increases, it increases the overall workload to the system for each customer request and essentially has the same effect as increasing the customer demand rate $\bar{\lambda}$. Consequently, the optimal price rate, the optimal wage rate and the optimal profit behave the same. This explains the third statement.

While optimal price rate is not necessarily monotonic with respect to K, μ and c, we can prove the following monotonicity property of the optimal payout ratio $\frac{w^*}{p^*}$ as as the model parameters change.

Proposition 2. *The optimal payout ratio* $\frac{w^*}{p^*}$ *increases in* c, $\bar{\lambda}$ *and* d, *and decreases in* K *and* μ.

Proposition 2 shows that the platform should increase the payout ratio $\frac{w^*}{p^*}$ to its providers when the customer's waiting cost c is higher, the maximum customer demand rate $\bar{\lambda}$ is higher or the average amount of service units d is higher. On the other hand, the platform should reduce the payout ratio when the maximum number of potential service providers K or the average service speed μ increases. One interesting implication of this result is that it would be more profitable for an on-demand transportation service platform (such as Uber) to increase the payout ratio to its participating drivers when the customer demand rate $\bar{\lambda}$ is higher and/or the travel speed μ is lower during rush hours.

Proposition 2 also indicates that it is more profitable for the platform to lower its payout ratio when the number of registered providers K increases. It is interesting to note that this result is consistent with Uber's strategy as reported by Huet (2014) that Uber offered a payout ratio of 0.8 for its first cohorts of drivers in San Francisco initially, but Uber lowered its payout ratio to 0.75 for its second cohorts of drivers in 2014 (i.e., as the number of registered drivers increases). Therefore, this result provides an economic justification for Uber to reduce its payout ratio as K increases.

3.4.1 Special case 1: when the payout ratio $\frac{w}{p}$ is fixed

As many on-demand service platforms (such as Uber and Didi) have adopted a fixed payout ratio to their service providers, we can adapt our base model to analyze this special case by imposing an additional constraint $\frac{w}{p} = \alpha$. We can use a similar analysis to establish the following result.

Proposition 3. *Under the additional constraint that* $\frac{w}{p} = \alpha$, $0 < \alpha < 1$, *both the optimal wage rate* w^* *and the optimal price rate* p^* *increase in* $\bar{\lambda}$ *and* d.

When the payout ratio is held constant so that $\frac{w}{p} = \alpha$, Proposition 3 implies that the optimal price rate p^* (and thus the optimal wage rate w^* due to a fixed payout ratio) should both be higher when customer demand rate for service $\bar{\lambda}$ is higher or when the average amount of service units d is higher. (We remark that the optimal price and wages rates, however, are not necessarily monotone in either the number of available service providers K, the average service speed μ, or the unit waiting cost c.) Our results thus suggest that an on-demand transportation service platform of using a fixed payout ratio (such as Uber) should charge a higher price (and thus provide a higher wage rate) during rush hours when the customer demand is high. This result is consistent with the notion of "surge pricing" as adopted by Uber and Lyft; see Cachon et al. (2015) for some recent discussions on the role of surge pricing.

We note that both Propositions 1 and 3 reveal that when the customer demand is higher, the platform should charge a higher price rate and offer a higher wage rate, regardless of whether the payout ratio is variable or fixed. In Section 6, we shall further compare the optimal profits of the platform between these two different settings using some numerical examples motivated by the sample data provided by Didi.

3.4.2 Special case 2: when the service level s is exogenously given

As on-demand service platforms continue to emerge and innovate, a new start-up platform might need to target a very high service level to ensure high customer satisfaction and gain popularity, at the expense of a lower near-term profit, during the initial phase of its operations. We can adapt our base model to analyze this special case by imposing a fixed target service level s. In other words, when the parameter s (or equivalently, the customer request rate λ because $\lambda = s\bar{\lambda}$) is exogenously given, the optimization problem of the platform is now reduced to:

$$\max_k \ \pi(k) \equiv \lambda d \left[(1-s) - c\left(\frac{\lambda d}{k\mu}\right) \frac{1}{k\mu - \lambda d} - \frac{k^2}{K\lambda d} \right], \text{ subject to } k > \frac{\lambda d}{\mu}.$$

It is straightforward to show that the above profit function $\pi(k)$ is concave and we can determine the optimal number of participating providers k^* using the first-order condition. Then, we can use (3.9) and (3.10) to retrieve the corresponding optimal price rate p^* and optimal wage rate w^* from the value of k^*. The following proposition summarizes the main results for this special case.[8]

Proposition 4. *The optimal price p^*, the optimal wage w^*, and the optimal profit π^* exhibit the following characteristics:*

1. *When K or μ increases, p^* increases, w^* decreases, and π^* increases.*
2. *When c increases, p^* decreases, w^* increases, and π^* decreases.*
3. *When $\bar{\lambda}$ or d increases, p^* decreases and w^* increases.*

Proposition 4 can be interpreted as follows. When the maximum number of potential providers K or the service speed μ becomes higher, the potential capacity of the system increases. The first statement asserts that it is then optimal for the platform to charge a higher price p^* (because of lower wait-time due to higher capacity), offer a lower wage w^* (because there are plenty of potential providers), and earn a higher profit π^*. The second statement states that when customers become less patient (i.e., when c increases), the platform should lower its price p^* (to compensate for the higher waiting cost), offer a higher wage w^* (to entice more providers to offer service), and consequently, the platform earns less. Finally, when the customer request rate λ (or equivalent, $\bar{\lambda}$, as $\lambda = s\bar{\lambda}$ and s is fixed) or the average amount of service unit d increases, the average workload of the system increases. As such, the third statement reveals that the platform should lower its price p^* to compensate for the higher waiting cost and offer a higher wage w^* to entice more providers to participate.

[8]The results of Proposition 4 continue to hold under more general distributions $F(.)$ or $G(.)$ and general wait-time function W_q. For ease of exposition, we shall relegate the details to Appendix A.)

By comparing the results of Propositions 1 and 4, one can observe that most of the results remain the same except for the characteristics of the optimal price rate p^*. When the service level s is endogenously determined, the optimal price rate p^* is not necessarily monotonic for certain model parameters, as stated in the first two statements of Proposition 1. However, when $\bar{\lambda}$ increases, the third statement of Proposition 1 reveals an opposite result, i.e., the optimal price rate p^* increases, versus p^* being decreasing in $\bar{\lambda}$ as given in Proposition 4. We can explain this opposite result as follows. When $\bar{\lambda}$ increases under a given s, the customer demand rate $\lambda = s\bar{\lambda}$ increases, and consequently, the platform has to offer a higher wage rate w^* to increase the number of providers k^*. Without the flexibility to adjust s, one can use (3.9) and the fact that $\lambda = s\bar{\lambda}$ to show that the corresponding optimal price rate p^* would decrease as stated in Proposition 4. On the other hand, when s (or thus the customer request rate λ) is endogenously determined, the platform has the flexibility to charge a higher price rate p^* and offer a higher wage rate w^* to better coordinate demand and supply as stated in Proposition 1.

We point out that Proposition 4 confirms the results (as also shown in Proposition 1) that it is always beneficial for the platform to recruit more potential service providers to join the platform (i.e., increase K), and to help (if possible) providers to increase their average service speed μ, regardless of whether the service level is fixed or endogenously determined.

We can use the results of Proposition 4 to characterize the optimal payout ratio $\frac{w^*}{p^*}$ as follows:

Corollary 1. *The optimal payout ratio $\frac{w^*}{p^*}$ increases in c, $\bar{\lambda}$ and d, but decreases in K and μ.*

The results in Corollary 1 are consistent with those given in Proposition 2. In particular, the platform should still increase the payout ratio $\frac{w^*}{p^*}$ to its providers when the customer's waiting cost c is higher or the maximum customer demand rate $\bar{\lambda}$ is higher, but should reduce the payout ratio when the maximum number of potential service providers K or the average service speed μ increases, even if the platform intends to maintain a constant target service level s at any time.

3.5. Extension: Social Welfare

Beside its own profit, the platform may have an interest in managing its customer and provider welfare carefully, especially when the practices of some service platforms could be potentially controversial. For example, the practice of Uber has been challenged by consumer rights group (out of concerns about public safety including sexual assaults, physical attacks) (e.g., Danielson (2015)), by independent drivers (due to their concerns about being treated as regular employees without benefits) (e.g., Roose (2014)), by the government (due to the concern over regulations), and by other taxi drivers (due to their concerns over unfair competition). In a law review article, Rogers (2015) provides a comprehensive list of social costs of Uber including: public safety, privacy, discrimination, labor law violations, etc. These social concerns have motivated us to extend our base model to incorporate social benefit. To do so, let us consider the case in which the platform expands the scope of its objective function by also including social (customer and provider) welfare in addition to its own profit.

In the same spirit as Cachon et al. (2015), we now extend our base model to the case when the objective is to maximize the total welfare, which includes both the customer and provider surpluses, and the firm's profit. First, for an individual customer who requests for service with a value rate of $v \geq F^{-1}(1-s)$, her surplus is given by $((v-p)d - cW_q)$. Therefore, the total customer surplus is equal to C_s, where

$$C_s = \bar{\lambda} \int_{F^{-1}(1-s)}^{1} [(v-p)d - cW_q] \, dF(v) = \bar{\lambda} \left[\left(\int_{F^{-1}(1-s)}^{1} v \; dF(v) - ps \right) d - cW_q s \right] \quad (3.12)$$

Similarly, for a provider who participates in the platform with a reservation rate of $r \leq w\frac{\lambda}{k\mu}$, his surplus is given by $w\frac{\lambda}{k\mu} - r$. Therefore, the total provider surplus is equal to P_s, where

$$P_s = K \int_0^{G^{-1}(\frac{k}{K})} \left(w\frac{\lambda d}{k} - r \right) dG(r) = w\lambda d - G^{-1}(\frac{k}{K})k + K \int_0^{G^{-1}(\frac{k}{K})} G(r)dr. \quad (3.13)$$

Then, the total welfare function can be expressed as $\Pi(k, s)$, where

$$\begin{aligned}\Pi(k, s) =& \pi(k, s) + \gamma(C_s + P_s) \\ =& \pi(k, s) + \gamma \left\{ \bar{\lambda} \left[\left(\int_{F^{-1}(1-s)}^1 v \, dF(v) - sp \right) d - scW_q \right] + w\lambda d - G^{-1}(\frac{k}{K})k + K \int_0^{G^{-1}(\frac{k}{K})} G(r)dr \right\} \\ =& \pi(k, s) + \gamma \left\{ \bar{\lambda} d \left[\int_{F^{-1}(1-s)}^1 v \, dF(v) - sF^{-1}(1-s) \right] + K \int_0^{G^{-1}(\frac{k}{K})} G(r)dr \right\}, \end{aligned} \quad (3.14)$$

where $\pi(k, s)$ is the profit function (3.7) for our base model, and $\gamma \in [0, 1]$ is the "equitable payoff" parameter which represents the willingness of the platform to give up some of its profit for a more equitable (or fairer) outcome for its customer and providers in setting its price and wage rates; see Cui et al. (2007). For example, when $\gamma = 1$, the platform weighs the social (customers and providers) welfare as equally important as its own profit. When $\gamma = 0$, the platform does not care about the social welfare so that $\Pi(k, s)$ simply reduces to $\pi(k, s)$ as given in (3.7).

In this extension, the platform's problem is to determine the optimal values of (k, s) that maximize the total welfare function $\Pi(k, s)$ subject to the system stability constraint: $k > \frac{\lambda d}{\mu}$. By using the same approach in analyzing the base model, we can obtain the following results:

Proposition 5. *When the platform is concerned about the total welfare as given in (3.14), the optimal solution exhibits the following characteristics:*

1. *All results as stated in Propositions 1,2,4 and Corollary 1 continue to hold.*

2. *When γ increases, the optimal wage rate w^* increases (and both s^* and k^* increase), but the optimal price rate p^* is not necessarily monotonic.*

Even when we incorporate both customer and provider surpluses in our objective function, the first statement of Proposition 5 shows that our analytical results for the base model as given in Section 4 are robust. The second statement shows that when a higher weight is placed on the social (customers and providers) welfare, the platform can increase the total welfare by increasing its target service level s^* (to serve more customers as $\lambda^* = s^*\bar{\lambda}$) and the wage rate w^* (to attract more providers k^* to participate). However, while both s^* and k^* increase in γ, Equation (3.9) reveals that the optimal price rate p^* is not necessarily monotonic as γ increases.

3.6. Numerical Illustrations Based on Didi Data

3.6.1 Background information

To illustrate the implications of our analytical results presented in this paper, we have collected real data from Didi, the largest on-demand ride sharing service platform operating in over 360 cities in China that was founded in June 2012.[9] Our data was based on rides that took place in the city of Hangzhou, the capitol city of Zhejiang province with an urban population of over 7 millions people, during the time periods between September 7-13 and November 1-30 in 2015. In Hangzhou city, Didi has approximately 13,000 registered drivers offering different types of services including Taxi (traditional taxi

[9]http://www.xiaojukeji.com/en/company.html. Didi merged with Kuaidi (a major competitor) in February 2015 as a way to defend its market share when Uber officially launched its service in China in July 2014. In August 2016, Uber decided to retreat from China and its China operations merged with Didi.

service),[10] Express/Private (equivalent to UBER X/Black with on-demand drivers), and Hitch (passengers sharing similar routes). For our numerical illustrations, we shall focus on the data associated with the Express/Private service, which accounts for 60% of all rides provided by Didi in Hangzhou. There were 13,000 registered drivers for all services, but the exact number of Express/Private drivers was not known to us. We shall assume that 60% of Didi drivers were Express/Private drivers, i.e., the number of registered Express/Private drivers in Hangzhou city was estimated to be $K = 7,800$.

3.6.2 Number of rides and drivers across different hours

Figure 1 depicts the average number of Express/Private rides and drivers across different hours on any given day. (Here, Hour 8 represents one-hour interval 8am-9am, Hour 19 for 7pm- 8pm, and so on. Data for Hours 1-7 were omitted due to incomplete data in the database.) We observe from the Didi data that the pattern depicted in Figure 2 is consistent throughout the weekdays, even though the average number of rides and drivers were slightly lower on Saturdays and Sundays, and that the peak hours are being Hours 9 and 19, and the slowest hours are being Hours 23 and 24. For instance, during the peak Hour 19, there were an average of 1,969 Express/Private rides and an average of 1,182 drivers in any given day. However, during the late night Hour 23, there were only an average of 1,033 rides and an average of 600 drivers.

3.6.3 Travel distance and travel speed

While the average number of rides and drivers vary substantially across different hours of the day, it is interesting to note from Figure 2 that the average travel distance for each

[10]Unlike Ubers business model that aims to displace the traditional taxi services, Didi integrates taxi services into its business model by providing its mobile hailing service to taxi drivers free of charge.

Express/Private ride is rather stable across different hours. For example, the average travel distance d during the peak Hour 19 and during the late night Hour 23 were 6.3 km and 6.6 km, respectively, while the average price per km p charged by Didi during these two hours were RMB 3.13 and RMB 2.76; respectively. We can also estimate the average travel speed across hours μ, and they were about 19 km/hour and 26 km/hour for Hour 19 and Hour 23, respectively. These numbers are thus consistent with the actual expected traffic conditions, which show that traffic is much less congested during late night hours.

3.6.4 Pricing and wage rates

Didi's price p for its service consists of two components so that $p = p_1 + p_2$, where p_1 represents the fare that is primarily based on the travel distance, and p_2 represents surcharges (e.g., tolls). Accordingly, Didi paid its drivers according to the following scheme. When a passenger pays a total fee of p, the driver will receive $(p_1 * 80\% - 0.5) * (100\% - 1.77\%) + p_2 * (100\% - 1.77\%)$ from Didi, but the driver needs to pay p_2 to cover various surcharges. Thus, the actual wage that Didi pays its drivers is approximately 80% of the total price; i.e., $w \approx 0.8p$. Figure 2 also shows that the average price per kilometer charged by Didi (excluding the surcharges) is relatively stable across different hours of the day. However, we observe that the average price per km p charged by Didi was RMB 3.13 during the peak hour (i.e., hour 19) and it was RMB 2.76 during the non-peak hour (i.e., hour 23). It is interesting to note that the observed price is higher during the peak hours, which corroborates with our results as stated in Proposition 3 that the optimal price rate p^* should be higher when customer demand rate for service is higher.

3.6.5 Strategic factors and their implications

It is important to note that the observed price that Didi charged its passengers was heavily discounted or subsidized during the data collection periods for the following two strategic reasons: (a) Didi wanted to attract more passengers by pricing its service below the traditional taxi services;[11] and (b) Didi was engaged in a price war to compete with Uber by offering discount coupons to compete for market share. In addition to offering heavily discounted price to attract passengers, Didi also provided extra "side payments" to its drivers to entice more drivers to join their platform due to the intense market competition. In addition to the regular payments of approximately 80% of the fare collected from the passengers, Didi (and Uber) had offered extra payments (e.g., Didi offers an extra bonus if the number of rides by a driver exceeds a certain quota within a 7-day period).

While we were unable to obtain the details of the bonus scheme, BBC (2016) had reported that the extra payment can be as high as 110% of the fare paid by the passengers. With such generous payments, more drivers reported to work and there was no need for Didi to use surge pricing to attract more drivers to offer rides during peak hours. This explains why Didi was able to offer relatively stable pricing in Hangzhou as depicted in Figure 2. Furthermore, the waiting time for Dids's service was reasonably short with an adequate supply of drivers. Specifically, the average waiting time of all Express/Private rides over the aforementioned time periods was about 6 minutes, of which the waiting time for accepting a ride request was approximately 1 minute, while the waiting time for picking up a passenger was approximately 5 minutes.

In view of the heavily discounted price due to the above strategic reasons, the price per km p as reported in Figure 2 was biased and did not represent the regular prices p and the actual

[11] In Hangzhou, taxi charges RMB 11 initially and then RMB 2.6 per km. As a way to entice passengers to choose Didi over taxi service, Didi had priced its service below taxi rates to increase market share. Based on our discussions with passengers in China, there was an expectation that Didi's price rate was lower than the taxi rate.

wages w the firm should offer in equilibrium. Nevertheless, we shall use the data given in the Didi database to construct our numerical examples to illustrate how our analytical results would compare with the actual prices/wages as reported in Didi's data set.

3.6.6 Numerical examples for illustrative purposes

We next provide some numerical examples based on the Didi data to illustrate our model results and discuss their implications. In all our numerical examples, we set the maximum number of drivers $K = 7,800$. We examined the average income for taxi drivers in Hangzhou city and the average major out-of-pocket expenses borne by the Didi drivers (including car insurance, license, fuel cost, etc.), and estimated that a minimum hourly wage of RMB 30 is required for a Didi driver to offer service. Thus, we assume that the hourly wage reservation r is distributed uniformly between RMB 30 to RMB 40 in our numerical examples.

As discussed earlier, the data were collected during the time when Didi (and Uber) was offering large fare discounts to attract riders, and so there was an expectation among riders that Didi price was less than the taxi rate in Hangzhou (which is RBM 2.6 per km). Thus, we used the taxi rate as a benchmark and assume that the customer value per km v is distributed uniformly between RMB 3 to RMB 4 in our numerical examples.

As shown in Figure 2, the average travel distances did not vary significantly across hours, so we simply set the average travel distance $d = 6$ km across all hours in order to focus our discussions on how different demand and congesting levels would affect the optimal price and wage rates across different hours of the day. It is difficult to provide an accurate estimate of the waiting cost parameter c, and so we simply varied the value of c from RMB 200 to RMB 2,200 per hour to illustrate how the optimal price and wage rates would change with respect to the cost of customer waiting for service.[12]

[12]While it is difficult to estimate the waiting cost, Gomez-Ibanez et al. (1999) reported that the waiting

We used data from two specific time periods to show our model results as illustrative examples for our discussions. In particular, we picked Hour 19 to represent peak-hour characteristics with high demand and travel congestion levels, and Hour 23 to represent non-peak hour characteristics with lower demand and congestion levels. Specifically, we set the average demand $\bar{\lambda} = 2,000$ with an average service speed $\mu = 19$ km/hour, and $\bar{\lambda} = 1,000$ with an average service speed $\mu = 26$ km/hour, respectively, in these two scenarios. In each scenario, we solved the base model as discussed in Section 4. The optimal number of participating drivers k^*, price rate p^* and wage rate w^* are given in Figures 3 and 4 for the peak hour and non-peak hour scenarios, respectively.

These numerical results illustrate the properties as stated in Proposition 1. For example, the optimal wages w^* increase as the waiting cost c increases in both Figures 3 and 4, which illustrates the results as stated in statement 2 of Proposition 1. By comparing the results in Figures 3 and 4, we can also observe that the values of k^* (scale on the left), p^* and w^* (scale on the right) are all higher during the peak hour (Figure 3) than those during the non-peak hour (Figure 4). These properties illustrate the results as stated in statements 1 and 3 of Proposition 1 (and Table 1), because the peak hour period has a higher customer demand rate $\bar{\lambda}$ and a slower service speed μ than that during non-peak hour period. However, the optimal number of participating driver k^* is not monotonic in the waiting cost c. In both scenarios, the optimal price rate p^* decreases as c increases. (However, the optimal price rate p^* is not necessarily monotonic in c (in general), as noted in statement 2 of Proposition 1.)

We also computed the optimal payout ratio $\alpha^* = \frac{w^*}{p^*}$; see Figure 5. The optimal payout ratio α^* increases from 0.68 to 0.84 for the peak hour scenario and from 0.54 to 0.72 for the

cost for a working class passenger in San Francisco is approximately 195% of the passenger's after-tax wages. Using this estimate and the fact that the average hourly wage of workers in Hangzhou is approximately RMB 40 per hour (China Daily, 2016), one can argue that the waiting cost for an average passenger in Hangzhou is approximately RMB 80 per hour. However, accounting for the income inequality and the impatient characteristics of most city dwellers in China (Li (2016)), we simply choose to consider the range of c varying from RMB 200 to RMB 2200 for illustrative purposes.

non-peak hour scenario, respectively, as c increases from 200 to 2,200. Also, observe from Figure 5 that the optimal payout ratio is always higher for the peak hour scenario than that for the non-peak hour scenario for any fixed value of c. (We note that these monotonicity properties are proved in Proposition 2.)

As Didi used a fixed payout ratio $\alpha \approx 0.8$, it would be interesting to compare the corresponding optimal profit between using the dynamic payout ratio α^* as given in our model versus using a fixed payout ratio $\alpha = 0.8$ to examine the potential benefits of adopting the optimal dynamic payout ratio. We illustrate our results in Figure 6 based on the peak hour scenario (i.e., Hour 19). Our numerical results show that, during these peak hour periods, using a dynamic payout ratio α^* can substantially increase the profit of the service platform over that using a fixed payout ratio of 0.8, especially when the waiting cost c is low when the optimal payout ratio is significantly different from 0.8 in our numerical examples here. For instance, when $c = 600$, the optimal profit is equal to 10,115 when the platform uses the optimal payout ratio $\alpha^* = 0.69$. However, if the platform uses a fixed payout ratio of 0.8, then the profit is equal to 7,001, which is much lower. However, it is important to point out that Didi (and Uber as well) used a fixed payout ratio due to various market considerations such as intense competition for driver participation and ridership as well as other practical implementation issues. Nevertheless, our results can serve as a guideline for understanding the magnitude of potential benefits for a hypothetical situation where such market considerations were no longer valid. Specifically, Figure 6 suggests that, when the waiting cost c is low, using a dynamic payout ratio can enable the platform to earn a much higher profit.

3.6.7 Additional observations from Didi data

We observe from the Didi data that the price per km p is highly correlated to the number of rides λ over the peak (non-peak) hours, with a correlation coefficient of 0.81. In other words, the price per km is usually higher during the time periods when the customer request rate is high (i.e., peak hours), and is lower during the time periods when the customer request rate is low (i.e., off-peak hours). Again, this pricing pattern corroborates with the results given in Proposition 3 that the optimal price rate p^* increases when the customer demand rate $\bar{\lambda}$ increases (while the average travel distance d is rather stable across different hours as noted in Section 6.3).

3.7. Conclusion

Motivated by the increasing popularity of on-demand service platforms with independent service providers and time-sensitive customers, we develop an analytical framework to understand how such platforms should set their optimal price and wage to match the needs of providers and customers taking into considerations the underlying supply and demand characteristics. The framework consists of a queueing model that captures some important market characteristics including wait-time sensitive customers and earnings sensitive suppliers. We analyze the steady state performance of a two sided queue in equilibrium and investigate the behavior of the optimal price and wage rates. We derive analytical results to show how different model parameters would affect these optimal price and wage rates. Our findings provide some interesting implications in managing prices and wages for on-demand service platforms.

Using some actual data collected from a major ride-sharing company in China, we construct some numerical examples to illustrate the results of our analytical model and discuss various implications on the optimal price and wage with respect to the underlying market characteristics. Although our model does not capture some important practical issues due to intense competition existed in China when the data were collected (and thus cannot be used to accurately predict the actual behavior of the players in the market), our analytical results can help to illustrate and explain some general observations that are consistent with the actual data provided by the company. More importantly, our model results can serve as a guideline for potentially increasing profitability when the underlying market conditions were to evolve to be consistent with the operating environment captured in our modeling framework. Specifically, we illustrate the potential benefits if the company were to adopt a dynamic payout ratio versus their current practice of using a fixed payout ratio.

Our model considers price and wage rates that are pre-committed, and we analyze the equilibrium behavior of the system. For future research, it would be interesting to study dynamic pricing strategies in which the platform can offer dynamic prices and wages to customers and providers based on the real-time status of the system. Specifically, one can develop a modeling framework that considers the real-time interactions among the customers, providers and the platform where the customers and providers need to make real-time decisions on whether to accept the dynamic prices and wages offered by the service platform.

3.8. Bibliography

[1] Afanasyev, M., H. Mendelson. (2010). Service provider competition: Delay cost structure, segmentation and cost advantage. *Manufacturing & Service Operation Management* 12(2): 213-235.

[2] Afeche, P., H. Mendelson. (2004). Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* 50(7): 869-882.

[3] Allon, G., A. Bassamboo, E.B. Cil. (2012). Large-scale service marketplaces: The role of the moderating firm. *Management Science* 58(10): 1854-1872.

[4] Anderson, M. (2006). Competition in two-sided markets. *The RAND Journal of Economics* 37(3): 668-691.

[5] Armony, M., and M. Haviv (2003). Price and delay competition between two service providers. *European Journal of Operational Research* 147(1): 32-50.

[6] Benjaafar, S., G. Kong, X. Li, and C. Courcoubetis. (2015). Peer-to-peer product sharing. Working paper, University of Minnesota.

[7] Cachon G.P., K.M. Daniels, R. Lobel. (2015) The role of surge pricing on a service platform with self-scheduling capacity. Available at SSRN.

[8] Chen, M.K., M. Sheldon. (2015) Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the UBER Platform. Working paper, UCLA Anderson School.

[9] Chen, X.M., M. Zahiri, S. Zhang. (2017). Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C* 76: 51-70.

[10] China Daily. 2016. Average salary in major Chinese cities is US$900 and growing. January 21, 2016. http://www.chinadaily.com.cn/china/2016-01/21/content_23183484.htm

[11] Cui, T.H., J.S. Raju, J. Zhang. (2007). Fairness and channel coordination. *management Science* 53(8): 1303-1314.

[12] Damodaran, A. (2014). A Disruptive Cab Ride to Riches: The UBER Payoff. *Forbes* (June 10) http://www.forbes.com/sites/aswathdamodaran/2014/06/10/a-disruptive-cab-ride-to-riches-the-uber-payoff/.

[13] Danielson, T., (2015) Uber just formed a board to address growing safety concerns. *Business Insiders*, November 24, 2015.

[14] DePhills, L. (2016). One Reason You Might be Better Off Driving for Uber than in a Taxi. *The Washington Post* (March 15).

[15] Fraiberger, S.P., A. Sundararajan. (2015). Peer-to-peer rental markets in the sharing economy. Working paper, New York University.

[16] Gomez-Ibanez, J., W. Tye, C. Winston. (1999). *Essays in Transportation Economics and Policy*, 42. Brookings Institution Press, Washington D.C.

[17] Gurvich, I., M. Lariviere, A. Moreno-Garcia. (2015). Operations in the on-demand economy: Staring services with self-scheduling capacity. Technical report, Northwestern University.

[18] Haws, K.L., X.O. Bearden. (2006). Dynamic pricing and consumer fairness perceptions. *Journal of Consumer Research* 33: 304-311.

[19] Hu, M., Y. Zhou. (2016) Take-Rate Crowdsourcing Contracts. Working paper. Rotman School of Management, University of Toronto.

[20] Huet, E. (2014) Uber Now Taking its Biggest UberX Commission Ever – 25 Percent. *Forbes*, September 22, 2014.

[21] Jiang, B., L. Tian. (2015). Collaborative consumption: Strategic and economic implications of product sharing. Working paper, Washington University.

[22] Klein, K. (2016) How This Startup Blew $13.5 Million and Ended in Bankruptcy. *Inc. Magazine* (April 5).

[23] Kokalitcheva, K. (2015). Uber and Lyft face a new challenger in Boston. *Fortune.com* (October 5).

[24] Lee, H.L., M.A. Cohen. (1983). A Note on the Convexity of Performance Measures of $M/M/c$ Queueing Systems. *Journal of Applied Probability* 20: 920-923.

[25] Li, A., (2016). Why are Chinese tourists so rude? A few insights. *South China Morning Post*, August 10, 2016.

[26] Li, J., A. Moreno, D.J. Zhang. (2015). Agent behavior in the sharing economy: Evidence from Airbnb. Working paper, University of Michigan .

[27] MacMillan, D. (2015). The $50 billion question: Can Uber deliver? *Wall Street Journal* (June 16) A1-A12.

[28] Moreno, A., C. Terwiesch. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research* 25(4): 865-886.

[29] Mosendz, P., H. Sender. (2014). Exclusive: Heres How Long It Takes to Get an Uber in U.S. Cities. *Tech & Science* (December 4).

[30] Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica* 37(1): 15-24.

[31] Riquelme, C., S. Banerjee, R. Johari. (2015). Pricing in ride-share platforms: A queueing-theoretic approach. Working paper, Stanford University.

[32] Rochet, J.C., J. Tirole. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association* 1(4): 990-1029.

[33] Rochet, J.C., J. Tirole. (2006). Two-sided markets: A progress report. *The RAND Journal of Economics* 37(3): 645-667

[34] Rogers, B. (2015). The Social Costs of Uber. *The University of Chicago Law Review*, 82: 85-104.

[35] Roose, K. (2014). Does Silicon Valley have a contract-worker problem? *NYMag.com* (September 18).

[36] Sakasegawa, H. (1997). An approximation formula $L_q \simeq \alpha\rho^\beta/(1-\rho)$. *Annals of the Institute of Statistical Mathematics* 29(1): 67-75.

[37] Sheldon, M. (2016). Income targeting and the ride-sharing market. Working paper, University of Chicago. Available at http://www.michaelsheldon.org/working-papers-section/.

[38] Shoot, B. (2015). Hot food, fast. *Entrepreneur* 68 (Aug.).

[39] Taylor T. (2016). On-Demand Service Platforms. Working paper, University of California, Berkeley. Available at SSRN 2722308.

[40] Wirtz, J., C.S. Tang. (2016). UBER: Competiting as Market Leader in the US versus Being a Distant Second in China. Case Study published in Wirtz and Lovelock (2016), Service Marketing: People, Technology and Strategy, 8th edition. World Scientific.

[41] Zhou, W., X. Chao, X. Gong. (2014). Optimal uniform pricing strategy of a service firm when facing two classes of customers. *Production and Operations Management* 23(4): 676-688.

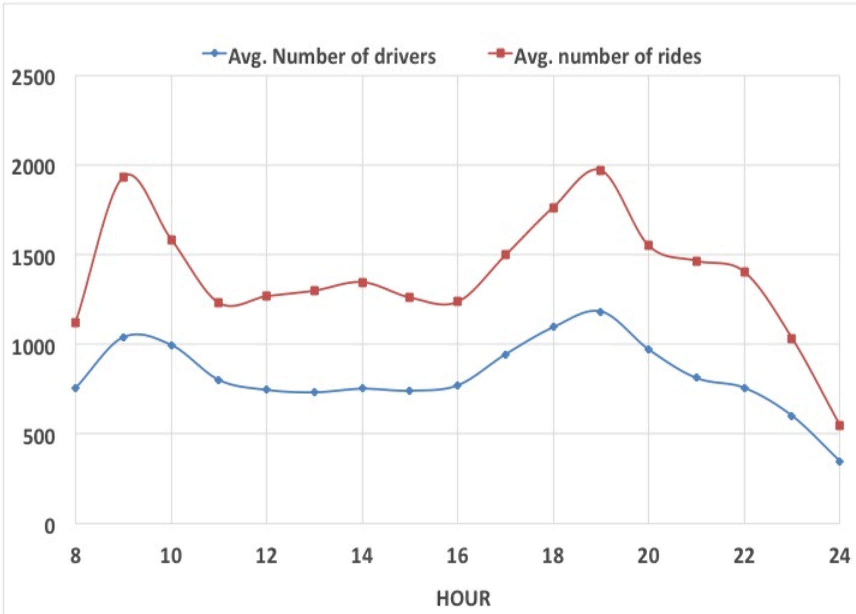


Figure 3.1: Number of rides and drivers across different hours.

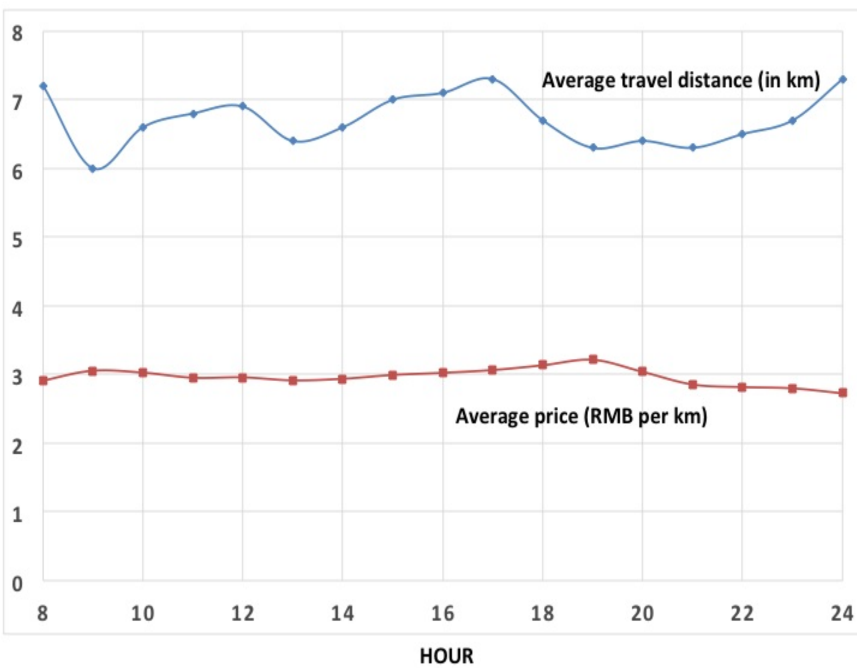


Figure 3.2: Average travel distance and average price per kilometer across different hours.

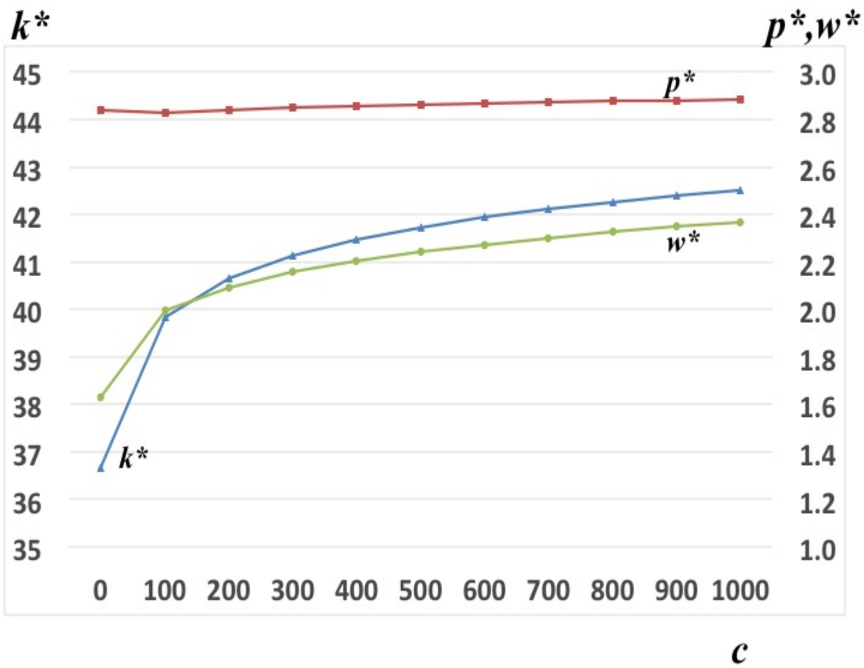


Figure 3.3: Optimal number of participating drivers, optimal price and wage rates during peak hours ($\bar{\lambda} = 2000$ and $\mu = 19$ km/hour).

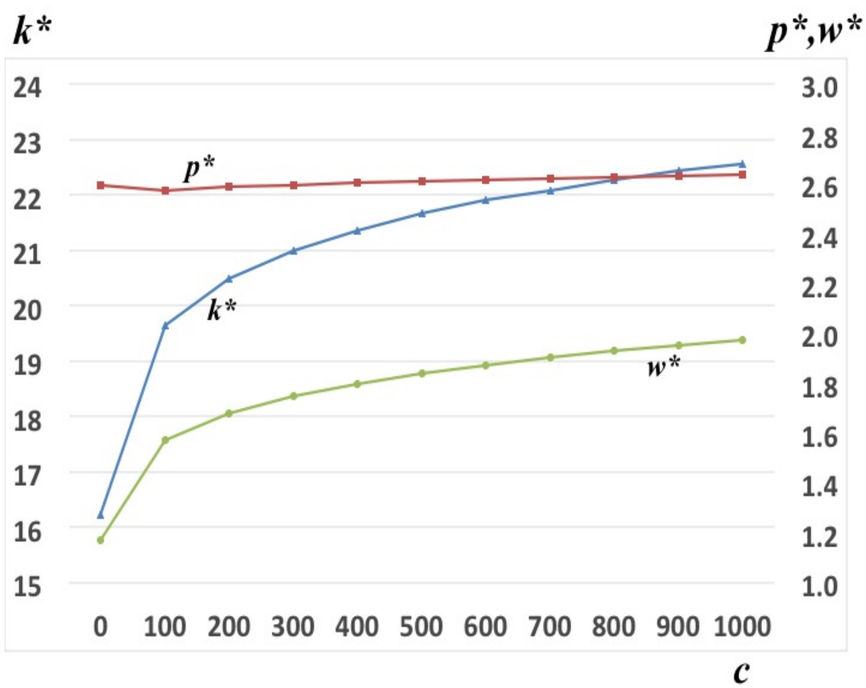


Figure 3.4: Optimal number of participating drivers, optimal price and wage rates during non-peak hours ($\bar{\lambda} = 1000$ and $\mu = 26$ km/hour).

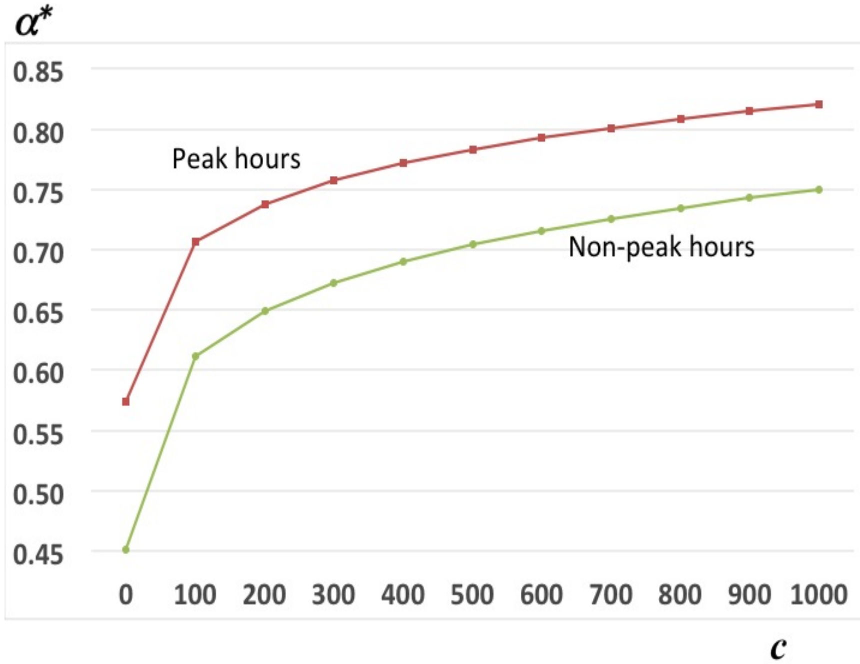


Figure 3.5: Comparisons of the optimal dynamic payout ratios between peak and non-peak hours.

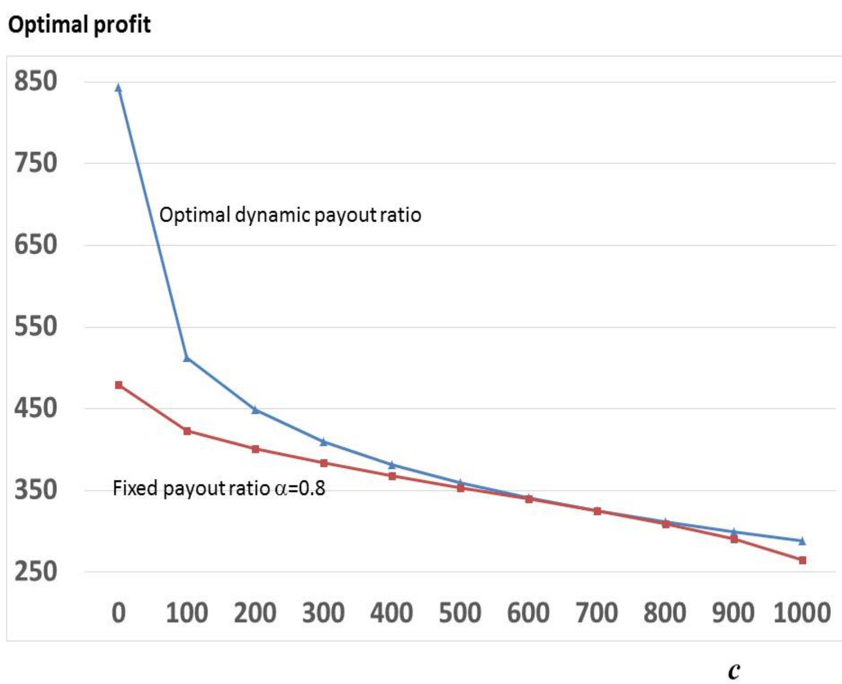


Figure 3.6: Comparisons of optimal profit between the optimal dynamic payout ratio and a fixed payout ratio for the peak hour scenario.

3.9. Appendix

Appendix A: Extensions of results in Proposition 4 under more general settings.

The results for our base model with exogenously given service level s as given in Section 4.2 are based on some simplifying assumptions as stated in Assumption 1. In this Appendix, we can relax these simplifying assumptions to more general settings in which W_q, $G(.)$ and $F(.)$ satisfy the following assumptions:

Assumption 2: *The wait-time function W_q is convex and increasing in λ, and is convex and decreasing in both k and μ. Furthermore, $\frac{\partial}{\partial \lambda}(\frac{\partial W_q}{\partial k}) < 0$, $\frac{\partial}{\partial d}(\frac{\partial W_q}{\partial k}) < 0$ and $\frac{\partial}{\partial \mu}(\frac{\partial W_q}{\partial k}) > 0$.*

Observe that the convexity of the waiting time function W_q is valid for an $M/M/k$ queueing model with arrival rate λ and service rate $\frac{\mu}{d}$; e.g., see Lee and Cohen (1983). The three conditions, $\frac{\partial}{\partial \lambda}(\frac{\partial W_q}{\partial k}) < 0$, $\frac{\partial}{\partial d}(\frac{\partial W_q}{\partial k}) < 0$ and $\frac{\partial}{\partial \mu}(\frac{\partial W_q}{\partial k}) > 0$, basically require that the marginal decrease in waiting time due to an additional service provider is larger at a higher system utilization level. This assumption is reasonable, and is also supported by the waiting time function of an $M/M/k$ queuing system. However, we do not require any specific functional form of W_q at this time.

Assumption 3: *The cumulative value distribution $F(.)$ is strictly increasing. The cumulative wage distribution $G(.)$ is concave and strictly increasing.*

Assumption 3 stipulates that the density of the reservation wage rate r is decreasing. This assumption implies that there are more service providers who would be willing to participate and offer service at a lower minimum earning rate.

By considering Assumptions 2 and 3 along with the profit function $\pi(k)$ given in (3.7), we obtain the following result:

Lemma 3.1. *The profit function $\pi(k)$ given in (3.7) is concave in k. Also, the optimal number of participating providers k^* satisfies the following first-order condition:*

$$-c\lambda \frac{\partial W_q}{\partial k} = G^{-1}(\frac{k}{K}) + G^{'-1}(\frac{k}{K})\frac{k}{K} = G^{-1}(\beta) + \beta G^{'-1}(\beta) = \frac{\partial(\beta G^{-1}(\beta))}{\partial \beta}. \tag{3.15}$$

Proof of Lemma 3.1: Differentiate the profit function given in (3.7) with respect to k and obtain

$$\pi'(k) = -c\lambda \frac{\partial W_q}{\partial k} - \left[G^{-1}(\frac{k}{K}) + G^{'-1}(\frac{k}{K})\frac{k}{K} \right] \tag{3.16}$$

and

$$\pi''(k) = -\lambda c \frac{\partial^2 W_q}{\partial k^2} - \left[2G^{'-1}(\frac{k}{K})\frac{1}{K} + G^{''-1}(\frac{k}{K})\frac{k}{K^2} \right]. \tag{3.17}$$

Assumption 3 implies that $G^{-1}(.)$ is convex and increasing. Together with Assumption 2, it follows that $\pi''(k) < 0$, which shows that $\pi(k)$ is concave in k. Therefore, the optimal value of k is given by the first-order condition $\pi'(k) = 0$, which is given in (3.15). This completes our proof. □

The first-order condition given in (3.15) can be interpreted as follows. The left side of (3.15) measures the marginal reduction in waiting cost for each additional service provider joining the platform. In view of (3.5), the term $G^{-1}(\beta) = w \cdot \frac{\lambda d}{k}$ represents the average earning rate of a provider. Hence, by noting that $\beta = k/K$, the right side of (3.15) can be interpreted as the marginal benefit associated with the increase in the average earning rate for each additional service provider participating in the platform in terms of β. Therefore, the first-order condition (3.15) shows that the optimal value of k is achieved when marginal cost equals marginal benefit.

By using the implicit function theorem to analyze the first-order condition (3.15), we can establish the following proposition in order to illustrate that some of the key results established in Proposition 4 continue to hold.

Proposition A1: Suppose that Assumptions 2 and 3 hold. Then,
(a) When K increases, both k^* and p^* increase, but the ratio $\beta = \frac{k^*}{K}$ decreases.
(b) When μ increases, both k^* and w^* decrease.
(c) When c increases, both k^* and w^* increase.
(d) When $\bar{\lambda}$ (or s) increases, k^* increases.
(e) When d increases, k^* increases.

Proof of Proposition A1: (a) Suppose that k_0 denotes the optimal value of k for $K = K_0$.

Using the first-order condition (3.15) and expressing the profit π as a function of k, we have

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K_0}) + G^{'-1}(\frac{k_0}{K_0})\frac{k_0}{K_0} \right] = 0 \tag{3.18}$$

Note that $G^{'-1}(.)$ is an increasing function since $G^{-1}(.)$ is convex as $G(.)$ is concave by Assumption 3, which implies that $\{G^{-1}(\frac{k_0}{K}) + G^{'-1}(\frac{k_0}{K})\frac{k_0}{K}\}$ is decreasing in K. Therefore, for any $K_1 > K_0$,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K_1}) + G^{'-1}(\frac{k_0}{K_1})\frac{k_0}{K_1} \right] > 0.$$

Since the profit function is concave in k, the optimal k^* must be greater than k_0 for any fixed $K = K_1 > K_0$, which shows that the optimal k^* is increasing in K. Since the waiting time W_q is decreasing in k, the optimal p^* given in (3.3) is also increasing in K.

Let $\beta_0 = \frac{k_0}{K_0}$, and rewrite the derivative of the profit function (3.18) as

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\beta_0) + G^{'-1}(\beta_0)\beta_0 \right\} = 0, \tag{3.19}$$

Let k_1 be the optimal value of k when $K = K_1 > K_0$, and define $\beta_1 = \frac{k_1}{K_1}$. Then,

$$\pi'(k_1) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_1} - \left\{ G^{-1}(\beta_1) + G^{'-1}(\beta_1)\beta_1 \right\} = 0. \tag{3.20}$$

Since k^* is increasing in K, we have $k_1 > k_0$. Thus, $-\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_1} < -\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0}$. From (3.19) and (3.20), we can obtain

$$G^{-1}(\beta_1) + G^{'-1}(\beta_1)\beta_1 < G^{-1}(\beta_0) + G^{'-1}(\beta_0)\beta_0.$$

Since $G^{-1}(\beta) + G^{'-1}(\beta)\beta$ is an increasing function in β, we can conclude that $\beta_1 < \beta_0$. Therefore, β^* is decreasing in K.

(b) Let k_0 denote the optimal k when $\mu = \mu_0$. Then,

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu_0, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G^{'-1}(\frac{k_0}{K})\frac{k_0}{K} \right] = 0$$

We have $\frac{\partial}{\partial \mu}(\frac{\partial W_q}{\partial k}) > 0$ from Assumption 2. Then, for any $\mu_1 > \mu_0$,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu_1, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] < 0,$$

Therefore, the optimal k^* must be smaller than k_0 for any fixed $\mu = \mu_1 > \mu_0$, which shows that the optimal k^* is decreasing in μ. From (3.6), the wage rate is increasing in k^*, therefore, w^* is decreasing in μ.

(c) Suppose that k_0 denotes the optimal value of k for $c = c_0$. Then,

$$\pi'(k_0) = -\lambda c_0 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] = 0$$

It is clear that for any $c_1 > c_0$,

$$-\lambda c_1 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] > 0.$$

Therefore, the optimal k^* must be greater than k_0 for any fixed $c = c_1 > c_0$, which shows that the optimal k^* is increasing in c. It is clear from (3.6) that the wage rate w is increasing in k. Therefore, the optimal w^* is also increasing in c.

(d) Since $\lambda = \bar{\lambda}s$, it suffices to show that k^* is increasing in λ. Let k_0 denote the optimal k when $\lambda = \lambda_0$. Then,

$$\pi'(k_0) = -\lambda_0 c \frac{\partial W_q(\lambda_0, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] = 0.$$

We have $\frac{\partial}{\partial \lambda}(\frac{\partial W_q}{\partial k}) < 0$ from Assumption 2, which implies that $-\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0}$ is increasing in λ. Therefore, for any $\lambda_1 > \lambda_0$, we have

$$-\lambda_1 c \frac{\partial W_q(\lambda_1, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] > 0.$$

Therefore, the optimal k^* must be greater than k_0 for any fixed $\lambda = \lambda_1 > \lambda_0$, which shows that the optimal k^* is increasing in λ.

(e) When $d = d_0$, let k_0 denote the optimal k. The first-order condition shows,

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d_0)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K} \right] = 0$$

Since from assumption 1, we know $\frac{\partial}{\partial d}(\frac{\partial W_q}{\partial k}) < 0$. Therefore, for any $d_1 > d_0$, we must have,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu, d_1)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K} \right] > 0,$$

The optimal k^* must be greater than k_0 for any $d = d_1 > d_0$, which shows that the optimal k^* is increasing in d. $\square$

Appendix B: Mathematical Proofs:

Proof of Proposition 1: It is straightforward to show that the profit function given in (3.7) is jointly concave over (s, k) under the assumptions that W_q is given by (3.8), $G(r) = r$ and $F(v) = v$. This implies that the optimal values of s^* and k^* are given by the two first-order conditions as given below:

$$\frac{\partial \pi}{\partial k} = cd^2\lambda \frac{\lambda(2k\mu - \lambda d)}{k^2\mu(k\mu - \lambda d)^2} - \frac{2k}{K} = 0 \tag{3.21}$$

$$\frac{\partial \pi}{\partial s} = \bar{\lambda} d \left\{ (1 - 2s) - \frac{c\lambda d(2k\mu - \lambda d)}{k\mu(k\mu - \lambda d)^2} \right\} = 0 \tag{3.22}$$

We shall use (3.22) to study the behavior of s^*, and (3.21) to characterize the behavior of k^*, as a function of s^*.

(a) Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $K = K_0$ and $K = K_1$, respectively. Suppose that $K_1 > K_0$. We shall show that $s_1 \geq s_0$ and $k_1 \geq k_0$, which implies that both s^* and k^* increase when K increases.

We use the notation $k^*(K, s)$ to denote the optimal value of k for the base model with parameter K and fixed service level s. In particular, $k^*(K_0, s_0) = k_0$ and $k^*(K_1, s_1) = k_1$. Since $K_1 > K_0$, it follows from Proposition A1(a) that $k^*(K_1, s_0) \geq k^*(K_0, s_0) = k_0$. It is clear that the derivative $\frac{\partial \pi}{\partial s}$ given in (3.22) is increasing in k. Since (s_0, k_0) satisfies the first-order condition $\frac{\partial \pi}{\partial s} = 0$ and $k^*(K_1, s_0) \geq k_0$, we have

$$(1 - 2s_0) - c\frac{\bar{\lambda} s_0 d \left[2k^*(K_1, s_0)\mu - \bar{\lambda} s_0 d\right]}{k^*(K_1, s_0)\mu \left[k^*(K_1, s_0)\mu - \bar{\lambda} s_0 d\right]^2} \geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $K = K_1$, as $\pi(s, k^*(s))$ is concave in s. Since (s_1, k_1) is optimal at $K = K_1$, this proves that $s_1 \geq s_0$. Also, it follows from Proposition A1(d) that $k_1 = k^*(K_1, s_1) \geq k^*(K_1, s_0) \geq k_0$. Therefore, we prove that both s^* and k^* increase in K.

Using (3.21) and (3.22), we have

$$1 - 2s = \frac{c\lambda d(2k\mu - \lambda d)}{k\mu(k\mu - \lambda d)^2} = 2w. \tag{3.23}$$

This proves that w^* decreases in K since s^* increases in K.

We next show that W_q^* decreases in K. First, we can rewrite (3.22) as

$$\bar{\lambda} sd(1 - 2s) = \lambda d \left[\frac{c\lambda d(2k\mu - \lambda d)}{k\mu(k\mu - \lambda d)^2}\right] = \frac{c\rho^2(2 - \rho)}{(1 - \rho)^2}. \tag{3.24}$$

Clearly, the right side of (3.24) is increasing in ρ. Also, the left side of (3.24) implies that $0 < s^* \leq \frac{1}{2}$. Suppose that $0 < s^* < \frac{1}{4}$. In this case, the left side of (3.24) increases in s. Since s^* increases in K as proved earlier, we can conclude that ρ^* must also increase in K in this case. The first-order condition (3.22) implies that

$$(1 - 2s^*) = \frac{cW_q^*(2k^*\mu - \lambda^*)}{d(k^*\mu - \lambda^*)} = \frac{c}{d} W_q^* \frac{2 - \rho^*}{1 - \rho^*}. \tag{3.25}$$

Since s^* increases in K, the left side of (3.25) must be decreasing as K increases. On the other hand, we have shown that ρ^* increases in K in this case, which implies that $\frac{2-\rho^*}{1-\rho^*}$ must be increasing in K in this case. We can conclude from (3.25) that W_q^* must be decreasing as K increases in this case.

Now suppose that $\frac{1}{4} \leq s^* \leq \frac{1}{2}$. In this case, the left side of (3.24) decreases in s. Since s^* increases in K, we can conclude from (3.24) that ρ^* must be decreasing in K in this case. The first-order condition (3.22) implies that

$$\bar{\lambda} s^*(1 - 2s^*) = \lambda^* \left[\frac{c\lambda^* d(2k^*\mu - \lambda^* d)}{k^*\mu(k^*\mu - \lambda^* d)^2}\right] = \frac{c}{d^3}(k^*\mu W_q^*)^2(2 - \rho^*). \tag{3.26}$$

As K increases, the left side of (3.26) decreases since s^* increases in K. On the right side of (3.26), both $(2 - \rho^*)$ and k^* increase with K as ρ^* increases in K in this case. Thus, we can conclude from (3.26) that W_q^* must also be decreasing in K in this case.

Furthermore,

$$\pi = \lambda d(p - w) = \lambda d(1 - s - \frac{c}{d}W_q - w) = \lambda d(\frac{1}{2} - \frac{c}{d}W_q), \tag{3.27}$$

where the last equality follows from (3.23). Since λ^* increases in K and W_q^* decreases in K, we can conclude that π^* increases in K.

(b) Similarly, let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $\mu = \mu_0$ and $\mu = \mu_1$, respectively. Suppose that $\mu_1 > \mu_0$. Again, we use the notation $k^*(\mu, s)$ to denote the optimal value of k for the base model with parameter μ and fixed service level s such that $k^*(\mu_0, s_0) = k_0$ and $k^*(\mu_1, s_1) = k_1$. We also use the notation $\rho^*(\mu, s)$ and $W_q^*(\mu, s)$ to denote the corresponding optimal values of ρ and W_q for the base model with fixed μ and s.

As will be proved in Proposition 4(b), $k^*(\mu, s)$, $\rho^*(\mu, s)$ and $W_q^*(\mu, s)$ all decrease in μ. Then, the function

$$H(\mu) = \frac{\lambda d\left[2k^*(\mu, s)\mu - \lambda d\right]}{k^*(\mu, s)\mu\left[k^*(\mu, s)\mu - \lambda d\right]^2} = \frac{W_q^*(\mu, s)}{d}\frac{2 - \rho^*(\mu, s)}{1 - \rho^*(\mu, s)}$$

decreases in μ.

Since (s_0, k_0) is the optimal solution when $\mu = \mu_0$, they must satisfy the first-order condition (3.22), i.e.,

$$\bar{\lambda} d\left\{(1 - 2s_0) - \frac{c\bar{\lambda}s_0 d\left[2k_0\mu_0 - \bar{\lambda}s_0 d\right]}{k_0\mu_0\left[k_0\mu_0 - \bar{\lambda}s_0 d\right]^2}\right\} = 0. \tag{3.28}$$

Since $H(\mu)$ decreases in μ and $\mu_1 > \mu_0$, it follows from (3.35) that

$$\bar{\lambda} d\left\{(1 - 2s_0) - \frac{c\bar{\lambda}s_0 d\left[2k^*(\mu_1, s)\mu_1 - \bar{\lambda}s_0 d\right]}{k^*(\mu_1, s)\mu_1\left[k^*(\mu_1, s)\mu_1 - \bar{\lambda}s_0 d\right]^2}\right\} \geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $\mu = \mu_1$, i.e., $s_1 \geq s_0$. This proves that s^* increases in μ. It then follows immediately from (3.23) that w^* decreases in μ.

Similarly, we then show that W_q^* decreases in μ. Suppose that $0 < s^* < \frac{1}{4}$. In (3.24), the left side increases in s and the right side is increasing in ρ. Since s^* increases in μ, therefore, ρ^*

must also increase in μ. In (3.25), the left side decreases in μ, as s^* is increasing in μ. Since ρ^* increases in μ in this case, $\frac{2-\rho^*}{1-\rho^*}$ must also increase in μ. We can conclude from (3.25) that W_q^* must be decreasing as μ.

On the other hand, suppose that $\frac{1}{4} \leq s^* \leq \frac{1}{2}$. In this case, the left side of (3.24) decreases in s. We just proved that s^* increases in μ, therefore, from (3.24) that ρ^* must be decreasing in μ in this case.

The first-order condition (3.21) implies that

$$c(W_q^*)^2(2-\rho^*) = \frac{2d^2}{\mu^2 K}$$

As μ increases, the right side decreases. On the left side, $(2-\rho^*)$ increases with μ as ρ^* increases in μ in this case. We can conclude that W_q^* must be decreasing in μ in this case.

Since $\lambda^* = \bar{\lambda} s^*$ increases and W_q^* decreases in μ, it follows from (3.27) that π^* increases in μ. This proves statement 1.

(c) Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $c = c_0$ and $c = c_1$, respectively. Suppose that $c_1 > c_0$. Here, we use the notation $k^*(c, s)$ to denote the optimal value of k for the base model with parameter c and fixed service level s such that $k^*(c_0, s_0) = k_0$ and $k^*(c_1, s_1) = k_1$.

Since $c_1 > c_0$, it follows from Proposition A1(b) that $k^*(c_1, s_0) \geq k^*(c_0, s_0) = k_0$. Then,

$$\frac{c_0 d^2 \lambda^2 (2k_0\mu - \lambda d)}{k_0 \mu (k_0 \mu - \lambda d)^2} = \frac{2k_0^2}{K} \leq \frac{2k^*(c_1, s_0)^2}{K} = \frac{c_1 d^2 \lambda^2 \left[2k^*(c_1, s_0)\mu - \lambda d\right]}{k^*(c_1, s_0)\mu \left[k^*(c_1, s_0)\mu - \lambda d\right]^2}, \tag{3.29}$$

where the two equalities come from the first-order condition (3.21) and the fact that k_0 and $k^*(c_1, s_0)$ are the optimal values of k for the base model with $s = s_0$ when $c = c_0$ and $c = c_1$, respectively.

Since (s_0, k_0) is the optimal solution when $c = c_0$, they must satisfy the first-order condition (3.22), i.e.,

$$(1 - 2s_0) - \frac{c_0 \lambda d (2k_0 \mu - \lambda d)}{k_0 \mu (k_0 \mu - \lambda d)^2} = 0. \tag{3.30}$$

Combining (3.29) and (3.30), we obtain

$$(1-2s_0) - \frac{c_1\lambda d\,[2k^*(c_1,s_0)\mu - \lambda d]}{k^*(c_1,s_0)\mu\,[k^*(c_1,s_0)\mu - \lambda d]^2} \leq 0.$$

Therefore, the optimal value of s must be smaller than s_0 when $c = c_1$, as $\pi(s, k^*(s))$ is concave in s. This proves that s^* is decreasing in c. It follows immediately from (3.23) that w^* is increasing in c.

Also, we can use (3.10) to express $\rho^* = \sqrt{\frac{\bar{\lambda}s^*d}{K\mu^2 w^*}}$, where $\rho^* = \frac{\lambda^* d}{k^*\mu}$ and $\lambda^* = \bar{\lambda}s^*$. Since s^* is decreasing in c and w^* is increasing in c, ρ^* is decreasing in c. Now rewrite (3.21) as

$$c(2-\rho)(\frac{\mu}{d}W_q)^2 = \frac{2}{K}, \tag{3.31}$$

which implies that W_q^* is decreasing in c since $2-\rho^*$ is increasing in c.

Since both s^* and ρ^* decrease in c, it follows from (3.25) that cW_q increases in c. Since $\lambda^* = \bar{\lambda}s^*$ decreases in c, it follows from (3.27) that π^* is decreasing in c. This proves statement 2.

(d) Now let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $\bar{\lambda} = \bar{\lambda}_0$ and $\bar{\lambda} = \bar{\lambda}_1$, respectively. Suppose that $\bar{\lambda}_1 > \bar{\lambda}_0$. Again, we use the notation $k^*(\bar{\lambda}, s)$ to denote the optimal value of k for the base model with parameter $\bar{\lambda}$ and fixed service level s. In particular, $k^*(\bar{\lambda}_0, s_0) = k_0$ and $k^*(\bar{\lambda}_1, s_1) = k_1$. We also use the notation $\rho^*(\bar{\lambda}, s)$ and $W_q^*(\bar{\lambda}, s)$ denote the corresponding optimal values of ρ and W_q for the base model with fixed $\bar{\lambda}$ and s.

As will be proved in Proposition 4(d), $k^*(\bar{\lambda}, s)$, $\rho^*(\bar{\lambda}, s)$ and $W_q^*(\bar{\lambda}, s)$ all increase in $\bar{\lambda}$. This implies that the function

$$H(\bar{\lambda}) = \frac{\lambda d^2[2k^*(\bar{\lambda}, s)\mu - \lambda d]}{k^*(\bar{\lambda}, s)\mu[k^*(\bar{\lambda}, s)\mu - \lambda d]^2} = W_q^*(\bar{\lambda}, s)\frac{2-\rho^*(\bar{\lambda}, s)}{1-\rho^*(\bar{\lambda}, s)}$$

increases in $\bar{\lambda}$ since $\frac{2-\rho}{1-\rho}$ is an increasing function in ρ.

Since (s_0, k_0) is the optimal solution when $\bar{\lambda} = \lambda_0$, they must satisfy the first-order condition

(3.22), i.e.,

$$\bar{\lambda}_0 d\left\{(1-2s_0) - \frac{c\bar{\lambda}_0 s_0 d[2k_0\mu - \bar{\lambda}_0 s_0 d]}{k_0\mu[k_0\mu - \bar{\lambda}_0 s_0 d]^2}\right\} = 0. \tag{3.32}$$

Since $H(\bar{\lambda})$ increases in $\bar{\lambda}$ and $\bar{\lambda}_1 > \bar{\lambda}_0$, it follows from (3.32) that

$$\bar{\lambda}_1 d\left\{(1-2s_0) - \frac{c\bar{\lambda}_1 s_0 d[2k^*(\bar{\lambda}_1, s_0)\mu - \bar{\lambda}_1 s_0 d]}{k^*(\bar{\lambda}_1, s_0)\mu[k^*(\bar{\lambda}_1, s_0)\mu - \bar{\lambda}_1 s_0 d]^2}\right\} \leq 0.$$

Therefore, the optimal value of s must be smaller than s_0 when $\bar{\lambda} = \bar{\lambda}_1$, i.e., $s_1 \leq s_0$. This proves that s^* decreases in $\bar{\lambda}$. Then, it follows immediately from (3.23) that w^* increases in $\bar{\lambda}$.

Using (3.31), $(2-\rho^*)$ and W_q^* must change in the opposite direction, which implies that ρ^* and W_q^* must change in the same direction as $\bar{\lambda}$ increases. Also we can rewrite (3.21) as

$$\frac{c\rho^2(2-\rho)}{(1-\rho)^2} = \frac{2k^2}{K}, \tag{3.33}$$

which implies that $\frac{\rho^{*2}(2-\rho^*)}{(1-\rho^*)^2}$ and k^* must change in the same direction. It is easy to show that $\frac{\rho^{*2}(2-\rho^*)}{(1-\rho^*)^2}$ is increasing in ρ^*, which implies that ρ^* and k^* must change in the same direction as $\bar{\lambda}$ increases. Thus, we can conclude that k^*, ρ^* and W_q^* must all change in the same direction when $\bar{\lambda}$ increases.

Since s^* decreases in $\bar{\lambda}$, the left side of (3.25) increases in $\bar{\lambda}$, which implies that the right side of (3.25) also increases in $\bar{\lambda}$. Since W_q^* and ρ^* must change in the same direction as $\bar{\lambda}$ increases, we can conclude that both W_q^* and ρ^* increases in $\bar{\lambda}$. Also, because k^*, ρ^* and W_q^* all change in the same direction when $\bar{\lambda}$ increases, we must have k^* increases in $\bar{\lambda}$. Also, we must have $\lambda^* = \frac{\rho^* k^* \mu}{d}$ increases in $\bar{\lambda}$.

Using (3.22) and (3.8), we obtain

$$s = \frac{1}{2}\left\{1 - \frac{c\lambda d(2k\mu - \lambda d)}{k\mu(k\mu - \lambda d)^2}\right\} = \frac{1}{2}\left\{1 - \frac{c}{d}W_q\frac{2-\rho}{1-\rho}\right\}.$$

We substitute the above equation into (3.9) to obtain

$$p = 1 - \frac{1}{2}\left\{1 - \frac{c}{d}W_q\frac{2-\rho}{1-\rho}\right\} - \frac{c}{d}W_q = \frac{1}{2}\left\{1 + \frac{c}{d}W_q\frac{\rho}{1-\rho}\right\}. \tag{3.34}$$

Since both W_q^* and ρ^* increases in $\bar{\lambda}$, it follows from (3.34) that p^* increases in $\bar{\lambda}$.

Let π_1^* and π_0^* denote the optimal profit when $\bar{\lambda} = \bar{\lambda}_1$ and $\bar{\lambda} = \bar{\lambda}_0$, respectively. Also, let $\pi^*(\bar{\lambda}, s)$ denote the optimal profit for the base model with fixed values of $\bar{\lambda}$ and s. For any $(\bar{\lambda}, s)$ with a fixed value of $\lambda = \bar{\lambda}s$ in the base model, observe from (3.44) that the optimal values of k remain the same. Furthermore, it follows from (3.8) and (3.10) that the corresponding values of W_q^* and w^* are also the same. Using (3.27), this implies that

$$\pi^*(\bar{\lambda}_1, \frac{\bar{\lambda}_0 s_0}{\bar{\lambda}_1}) = \bar{\lambda}_0 s_0 d(1 - \frac{\bar{\lambda}_0 s_0}{\bar{\lambda}_1} - \frac{c}{d}W_q^* - w^*) \geq \bar{\lambda}_0 s_0 d(1 - s_0 - \frac{c}{d}W_q^* - w^*) = \pi^*(\bar{\lambda}_0, s_0),$$

when $\bar{\lambda}_1 > \bar{\lambda}_0$. Then,

$$\pi_1^* = \pi^*(\bar{\lambda}_1, s_1) \geq \pi^*(\bar{\lambda}_1, \frac{\bar{\lambda}_0 s_0}{\bar{\lambda}_1}) \geq \pi^*(\bar{\lambda}_0, s_0) = \pi_0^*,$$

where the first inequality is due to the fact that (k_1, s_1) is the optimal solution when $\bar{\lambda} = \bar{\lambda}_1$. This proves that π^* increases in $\bar{\lambda}$.

(e) Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $d = d_0$ and $d = d_1 > d_0$, respectively. We use the notation $k^*(d, s)$ to denote the optimal value of k for the base model with parameter d and fixed service level s such that $k^*(d_0, s_0) = k_0$ and $k^*(d_1, s_1) = k_1$. We also use the notation $\rho^*(d, s)$ and $W_q^*(d, s)$ to denote the corresponding optimal values of ρ and W_q for the base model with fixed d and s.

As will be proved in Proposition 4(e), $k^*(d, s)$, $\rho^*(d, s)$ and $\frac{W_q^*(d,s)}{d}$ all increase when d increases. Then,

$$H(d) = \frac{\lambda d\left[2k^*(d, s)\mu - \lambda d\right]}{k^*(d, s)\mu\left[k^*(d, s)\mu - \lambda d\right]^2} = \frac{W_q^*(d, s)}{d}\frac{2 - \rho^*(d, s)}{1 - \rho^*(d, s)}$$

increases in d.

Since (s_0, k_0) is the optimal solution when $d = d_0$, they must satisfy the first-order condition

(3.22), i.e.,

$$\bar{\lambda} d_0 \left\{ (1-2s_0) - \frac{c\bar{\lambda} s_0 d_0 \left[2k_0\mu - \bar{\lambda} s_0 d_0\right]}{k_0 \mu \left[k_0\mu - \bar{\lambda} s_0 d_0\right]^2} \right\} = 0. \tag{3.35}$$

Since $H(d)$ increases in d and $d_1 > d_0$, it follows from (3.35) that

$$\bar{\lambda} d_1 \left\{ (1-2s_0) - \frac{c\bar{\lambda} s_0 d_1 \left[2k^*(d_1, s_0)\mu - \bar{\lambda} s_0 d_1\right]}{k^*(d_1, s_0)\mu \left[k^*(d_1, s_0)\mu - \bar{\lambda} s_0 d_1\right]^2} \right\} \leq 0.$$

Therefore, the optimal value of s must be smaller than s_0 when $d = d_1$, i.e., $s_1 \leq s_0$. This proves that s^* decreases in d. It then follows immediately from (3.23) that w^* increases in d.

We can use (3.31) to deduce that ρ^* and $\frac{W_q^*}{d}$ must change in the same direction when d increases. We can also use (3.33) to deduce that ρ^* and k^* must change in the same direction when d increases. Thus, we can conclude that k^*, ρ^* and $\frac{W_q^*}{d}$ must all change in the same direction when d increases. Since s^* decreases in d, we can use (3.25) and the fact that both ρ^* and $\frac{W_q^*}{d}$ must change in the same direction to conclude that both ρ^* and $\frac{W_q^*}{d}$ increase in d, which also implies that k^* and W_q^* increase in d. It also follows from (3.34) that p^* increases in d.

Let π_1^* and π_0^* denote the optimal profit when $d = d_1$ and $d = d_0$, respectively. Also, let $\pi^*(d, s)$ denote the optimal profit for the base model with any fixed values of d and s. For any (d, s) with a fixed ratio of ds in the base model, it is easy to check from (3.44) that the optimal values of k remain the same, and from (3.8) and (3.10) that the corresponding values of $\tilde{W}_q^* = \frac{W_q^*}{d}$ and w^* are also the same. Then,

$$\pi^*(d_1, \frac{d_0 s_0}{d_1}) = \bar{\lambda} s_0 d_0 (1 - \frac{d_0 s_0}{d_1} - c\tilde{W}_q^* - w^*) \geq \bar{\lambda} s_0 d_0 (1 - s_0 - c\tilde{W}_q^* - w^*) = \pi^*(d_0, s_0),$$

when $d_1 > d_0$. Then,

$$\pi_1^* = \pi^*(d_1, s_1) \geq \pi^*(d_1, \frac{d_0 s_0}{d_1}) \geq \pi^*(d_0, s_0) = \pi_0^*,$$

where the first inequality is due to the fact that (k_1, s_1) is the optimal solution when $d = d_1$. Therefore, π^* increases in d. This proves statement 3. □

Proof of Proposition 2: Let $\alpha^* = \frac{w^*}{p^*}$. As shown in (3.27), we can express

$$p^* - w^* = (\frac{1}{\alpha^*} - 1)w^* = \frac{1}{2} - \frac{c}{d}W_q^*. \tag{3.36}$$

We have shown in the proof of Proposition 1(c) that w^* and cW_q^* increase in c and in Proposition 1(d) and (e) that w^* and $\frac{W_q^*}{d}$ increase in $\bar{\lambda}$ and d. We can then conclude from (3.36) that α^* is increasing in c, $\bar{\lambda}$ and d. On the other hand, we have shown in the proof of Proposition 1(a) that w^* and W_q^* decrease in K and Proposition 1(b) that w^* and W_q^* decrease in μ. Again, we can conclude from (3.36) that α^* is decreasing in K and μ. □

Proof of Proposition 3: With the constraint that $\frac{w}{p} = \alpha$, the objective function can be expressed as $\pi = \lambda d(p - w) = \lambda d(\frac{1}{\alpha} - 1)w$. We can solve the constrained problem as an unconstrained Lagrange optimization problem with the Lagrange function of $L(p, w, z) = \lambda d(\frac{1}{\alpha} - 1)w + z\lambda d(\alpha p - w)$, where z is the nonzero Lagrange multiplier.[13] We substitute the values of p and w given by (3.9) and (3.10) and the fact that $\lambda = \bar{\lambda}s$ into the Lagrange function $L(p, w, z)$, and can obtain the following optimality conditions from the three first-order conditions, $\frac{\delta L}{\delta k} = 0$, $\frac{\delta L}{\delta s} = 0$, and $\frac{\delta L}{\delta z} = 0$, respectively:

$$(\frac{1}{\alpha} - 1)\frac{2k}{K} + z\left[\frac{\alpha c\lambda^2 d^2(2k\mu - \lambda d)}{k^2\mu(k\mu - \lambda d)^2} - \frac{2k}{K}\right] = 0, \tag{3.37}$$

$$(1 - 2s) - \frac{c\lambda d(2k\mu - \lambda d)}{k\mu(k\mu - \lambda d)^2} = 0, \tag{3.38}$$

$$\alpha\left[1 - s - \frac{c\lambda d}{k\mu(k\mu - \lambda d)}\right] - \frac{k^2}{K\lambda d} = 0. \tag{3.39}$$

We next use the optimality conditions (3.38) and (3.39) to establish the following properties:
(i) λ^* and k^* change in the same direction for any fixed α, K, c, d and μ;
(ii) ρ^* and k^* change in the same direction for any fixed α, K, c and μ;
(iii) ρ^* and w^* change in the same direction for any fixed α, K, c and μ;
(iv) ρ^* and $\frac{W_q^*}{d}$ change in the same direction for any fixed α, K, c and μ;
(v) ρ^* and s^* change in the opposite direction for any fixed α, K and c.

[13] We ignore the constraints that $0 \le s \le 1$ and $0 \le k \le K$ to simplify our exposition in the proof, but the analysis can be easily adapted to include these constraints as well.

First, we can substitute (3.38) into (3.39) and obtain

$$\frac{\alpha}{2}\left[1+\frac{c\lambda^2 d^2}{k\mu(k\mu-\lambda d)^2}\right]-\frac{k^2}{K\lambda d}=0. \tag{3.40}$$

The left side of (3.40) increases in λ but decreases in k for any fixed α, K, c, d and μ. Thus, the values of λ and k at optimality must change in the same direction, which proves (i).

Let $\rho = \frac{\lambda d}{k\mu} < 1$. We can rewrite (3.40) as

$$\frac{\alpha}{2}\left[\frac{\rho\mu}{k}+\frac{c\rho^3}{k^2(1-\rho)^2}\right]-\frac{1}{K}=0. \tag{3.41}$$

The left side of (3.41) increases in ρ but decreases in k for any fixed α, K, c and μ. Thus, the values of ρ and k at optimality must change in the same direction, which prove (ii).

We can use (3.10) to rewrite (3.40) as

$$\frac{\alpha}{2}\left[1+\frac{c\rho}{(1-\rho)^2 K\mu^2 w}\right]-w=0. \tag{3.42}$$

The left side of (3.42) increases in ρ but decreases in w for any fixed α, K, c and μ. Thus, the values of ρ and w at optimality must change in the same direction, which proves (iii).

We can use (3.8) to rewrite (3.40) as

$$\rho=\frac{1-(2d)/(K\alpha\mu^2 W_q)}{1-\frac{c}{d}W_q}.$$

This shows that the values of $\frac{W_q}{d}$ and ρ at optimality must change in the same direction for any fixed α, K, c and μ, which proves (iv).

Finally, we can again use (3.8) to rewrite (3.38) as

$$(1-2s)-\frac{c}{d}W_q\frac{2-\rho}{1-\rho}=0,$$

or equivalently,

$$s = \frac{1}{2}\left[1 - \frac{c}{d}W_q\frac{2-\rho}{1-\rho}\right].$$

We show in (iv) that the values of ρ and $\frac{W_q}{d}$ at optimality change in the same direction for any fixed α, K and c. It then follows that the value of ρ and s at optimality must change in the opposite direction, which proves (v).

We can now rewrite (3.38) as

$$(1 - 2\frac{\lambda}{\bar{\lambda}}) - \frac{c}{d}W_q\frac{2-\rho}{1-\rho} = 0. \tag{3.43}$$

We have shown in (i), (ii) and (iv) that the values of λ, $\frac{W_q}{d}$ and ρ at optimality change in the same direction for any fixed α, K, c, d and μ. We can deduce from (3.43) that the values of λ, $\frac{W_q}{d}$ and ρ at optimality must all increase when $\bar{\lambda}$ increases. It then follow from (iii) that w^* (and thus $p^* = w^*/\alpha$) increases when $\bar{\lambda}$ increases.

We can also use $\rho = \frac{\bar{\lambda}s^*d}{k\mu}$ to express d as $d = \frac{\rho k\mu}{\bar{\lambda}s}$. It follows from (ii) and (v) that the value of ρ at optimality must change in the same direction as k, but in opposite direction of s for any fixed α, K, c and μ. Therefore, we can deduce that, at optimality, the values of ρ and k must increase while the value of s must decrease when d increases. From (iii) we can also conclude that w^* (and thus $p^* = w^*/\alpha$) increases when d increases. $\square$

Proof of Proposition 4: We shall prove the results for more general distributions of $F(.)$ and $G(.)$ satisfying Assumption 3, which clearly includes the uniform [0,1] distribution. Note that the wait-time function W_q given in (3.8) satisfies Assumption 2, so that the results of Proposition A1 hold.

(a) Since Proposition A1(a) holds, we know that p^* increases as K increases. We need to show that w^* decreases as K increases. We can differentiate the profit function π given in (3.11) with respect to k and obtain the first-order condition

$$\pi'(k) = cd^2\lambda\frac{\lambda(2k\mu - \lambda d)}{k^2\mu(k\mu - \lambda d)^2} - \frac{2k}{K} = 0. \tag{3.44}$$

Using (3.10) and (3.44), we can express

$$w = c\lambda d \frac{2k\mu - \lambda d}{2k\mu(k\mu - \lambda d)^2}.$$

Therefore,

$$\frac{\partial w}{\partial k} = -c\lambda d \frac{3k\mu(k\mu - \lambda d) + \lambda^2 d^2 + k^2\mu^2}{2k^2\mu(k\mu - \lambda d)^3} < 0,$$

which implies that w^* decreases as K increases, as k^* is increasing in K from Proposition A1(a). Since $\pi^* = \lambda d(p^* - w^*)$, π^* increases as K increases.

(b) We can use the same argument as given in part (d) and Proposition A1(b) to show that w^*, $\frac{\mu}{d}W_q^*$ and ρ^* decrease when μ increases. It follows that $\frac{1}{d}W_q^*$ must be decreasing when μ increases. Therefore, from (3.3), p^* increases when μ increases. Since $\pi^* = \lambda d(p^* - w^*)$, π^* increases as μ increases. This proves statement 1.

(c) Consider any $c_1 > c_0 > 0$, and let k_i be the corresponding optimal value of k^* when $c = c_i$, $i = 1, 2$. Then, $k_0 < k_1$, as k^* is increasing in c from Proposition A1(c). Also,

$$\frac{\frac{\partial W_q}{\partial k}}{W_q} = \frac{\lambda d - 2k\mu}{k(k\mu - \lambda d)} = -\frac{1}{k} - \frac{\mu}{k\mu - \lambda d},$$

is increasing in k. Since $k_0 < k_1$, we have

$$\frac{\frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0}}{W_q(\lambda, k_0, \mu, d)} < \frac{\frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_1}}{W_q(\lambda, k_1, \mu, d)}. \tag{3.45}$$

By the definition of k_i and using the first-order condition as in the proof of Proposition A1(c), we have

$$-\lambda c_0 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] = 0$$

and

$$-\lambda c_1 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_1} - \left[G^{-1}(\frac{k_1}{K}) + G'^{-1}(\frac{k_1}{K})\frac{k_1}{K}\right] = 0.$$

Since $k_0 < k_1$ and $G^{-1}(.)$ is convex and increasing, the above two equations imply that

$$-c_0 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} < -c_1 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_1}. \tag{3.46}$$

Using (3.45) and (3.46), we can conclude that $c_0 W_q(\lambda, k_0, \mu, d) < c_1 W_q(\lambda, k_1, \mu, d)$. It then follows directly from (3.3) that the optimal price when $c = c_0$ is higher than the optimal price when $c = c_1$. This shows that p^* is decreasing in c. It follows from Proposition $A1(c)$ that w^* is increasing in c, as Assumptions 2 and 3 hold. Also, $\pi^* = \lambda d(p^* - w^*)$ is decreasing in c. This proves statement 2.

(d) Since $\lambda = \bar{\lambda} s$, it suffices to show that p^* is decreasing in λ and w^* is increasing in λ. Let $\rho^* = \frac{\lambda d}{k^* \mu}$ denote the provider utilization at optimality, so that we can rewrite the first-order condition (3.44) as

$$\frac{c(\rho^*)^2(2-\rho^*)}{(1-\rho^*)^2} = \frac{2(k^*)^2}{K},$$

or equivalently,

$$k^* = \sqrt{\frac{cK}{2(2-\rho^*)} \frac{\rho^*(2-\rho^*)}{1-\rho^*}}.$$

It is clear from the above equation that ρ^* increases as k^* increases. It then follows from Proposition A1(d) that ρ^* increases when $\bar{\lambda}$ or s increases.

Also, denote the waiting time at optimality, $W_q^* = \frac{\lambda^* d^2}{k^* \mu (k^* \mu - \lambda^* d)}$, so that we can rewrite the first-order condition (3.44) as

$$c(\frac{\mu}{d} W_q^*)^2 (2-\rho^*) = \frac{2}{K}.$$

This implies that $\frac{\mu}{d} W_q^*$ increases as ρ^* increases. It then follows from (3.3) that p^* decreases when $\bar{\lambda}$ or s increases.

Finally, since $w^* = \frac{(k^*)^2}{K \lambda d}$, we can also rewrite (3.44) as

$$w^* = \frac{c \frac{\mu}{d} W_q^*}{2\mu} \left(\frac{2-\rho^*}{1-\rho^*} \right). \tag{3.47}$$

Since both $\frac{\mu}{d}W_q^*$ and ρ^* increase when $\bar{\lambda}$ or s increases, we prove the result that w^* increases when $\bar{\lambda}$ or s increases.

(e) We can use the same argument as given in part (d) and Proposition A1(e) to show that p^* decreases and w^* increases when d increases. This proves statement 3. $\square$

Proof of Proposition 5: (a) Let us first adapt the proof of Proposition A1 to establish the same results to this extension. To establish the result of Proposition A1(a), let k_0 denotes the optimal value of k for $K = K_0$. The first-order condition for $\Pi(k)$ now becomes

$$\Pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\frac{k_0}{K_0}) + (1-\gamma)G'^{-1}(\frac{k_0}{K_0})\frac{k_0}{K_0} \right\} = 0, \quad (3.48)$$

and we can show that, for any $K_1 > K_0$,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\frac{k_0}{K_1}) + (1-\gamma)G'^{-1}(\frac{k_0}{K_1})\frac{k_0}{K_1} \right\} > 0.$$

Therefore, the optimal value of k must be greater than k_0 for any fixed $K_1 > K_0$, which implies that k^* is increasing in K. Using the same argument as before, we can show that p^* is increasing in K.

Let $\beta_0 = \frac{k_0}{K_0}$, we can rewrite the first-order condition (3.48) as

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\beta_0) + (1-\gamma)G'^{-1}(\beta_0)\beta_0 \right\} = 0, \quad (3.49)$$

Let k_1 denote the optimal value of k when $K = K_1 > K_0$, and define $\beta_1 = \frac{k_1}{K_1}$. Then,

$$\pi'(k_1) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_1} - \left\{ G^{-1}(\beta_1) + (1-\gamma)G'^{-1}(\beta_1)\beta_1 \right\} = 0. \quad (3.50)$$

As k^* is increasing in K, we have $k_1 > k_0$. Therefore, $-\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_1} < -\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0}$. It follows from (3.49) and (3.50) that

$$G^{-1}(\beta_1) + (1-\gamma)G'^{-1}(\beta_1)\beta_1 < G^{-1}(\beta_0) + (1-\gamma)G'^{-1}(\beta_0)\beta_0.$$

Since $G^{-1}(\beta) + (1-\gamma)G'^{-1}(\beta)\beta$ is an increasing function in β, we can conclude that $\beta_1 < \beta_0$.

Thus, β^* is decreasing in K.

The results for Proposition A1 (b),(c),(d) and (e) can be proved using the same arguments from the proof of Proposition A1.

We can then easily adapt the proof of Proposition 4 to establish the same results to this extension. In this case, the first-order condition (3.44) given in the proof of Proposition 4 now becomes

$$\Pi'(k) = c\lambda \frac{\lambda d^2(2k\mu - \lambda d)}{k^2\mu(k\mu - \lambda d)^2} - (2-\gamma)\frac{k}{K} = 0, \tag{3.51}$$

and we can use the same arguments as before to establish the results for Proposition 4 and Corollary 1 in this extension. We omit the details here.

We next adapt the proofs of Propositions 1 and 2 to establish all the corresponding results to this extension. To illustrate the adaptation, we next outline the proof of the results of Proposition 1(a). The rest of the results can be proved by following the same arguments, the details will be omitted here.

The two first-order conditions (3.21) and (3.22) given in the proof of Proposition 1 now become

$$\frac{\partial \pi}{\partial k} = \frac{c\lambda^2 d^2(2k\mu - \lambda d)}{k^2\mu(k\mu - \lambda d)^2} - \frac{(2-\gamma)k}{K} = 0 \tag{3.52}$$

$$\frac{\partial \pi}{\partial s} = \bar{\lambda} d\left\{[1-(2-\gamma)s] - \frac{c\lambda d(2k\mu - \lambda d)}{k\mu(k\mu - \lambda d)^2}\right\} = 0. \tag{3.53}$$

Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $K = K_0$ and $K = K_1$, respectively. Suppose that $K_1 > K_0$. We use the notation $k^*(K, s)$ to denote the optimal value of k for the base model with parameter K and fixed service level s. Then, $k^*(K_1, s_0) > k^*(K_0, s_0)$ in view of Proposition A1(a). Furthermore, we can use the first-order condition to establish that

$$[1-(2-\gamma)s_0] - c\frac{\bar{\lambda}s_0 d\left[2k^*(K_1, s_0)\mu - \bar{\lambda}s_0 d\right]}{k^*(K_1, s_0)\mu\left[k^*(K_1, s_0)\mu - \bar{\lambda}s_0 d\right]^2} \geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $K = K_1$, as $\pi(s, k^*(s))$ is concave in s. Since (s_1, k_1) is optimal at $K = K_1$, this proves that $s_1 \geq s_0$. Also, it follows from Proposition A1(d) that $k_1 = k^*(K_1, s_1) \geq k^*(K_1, s_0) \geq k_0$. Therefore, we prove that both s^* and k^* increase in K.

Using (3.52) and (3.53), we can obtain

$$s + w = \frac{1}{2-\gamma}.$$

It then follows that w^* decreases in K, as s^* increases in K.

We next show that W_q^* decreases in K. We can rewrite (3.53) as

$$\bar{\lambda} sd[1 - (2-\gamma)s] = \left[\frac{c\rho^2(2-\rho)}{(1-\rho)^2}\right]. \tag{3.54}$$

Clearly, the right side of (3.54) is increasing in ρ. Also, the left side of (3.54) implies that $0 < s^* \leq \frac{1}{2-\gamma}$. Suppose that $0 < s^* < \frac{1}{2(2-\gamma)}$. In this case, the left side of (3.54) increases in s. Since s^* increases in K as proved earlier, we can conclude that ρ^* must also increase in K in this case. The first-order condition (3.53) implies that

$$[1 - (2-\gamma)s^*] = \frac{cW_q^*(2k^*\mu - \lambda^*)}{d(k^*\mu - \lambda^*)} = \frac{c}{d} W_q^* \frac{2-\rho^*}{1-\rho^*}. \tag{3.55}$$

Since s^* increases in K, the left side of (3.55) must be decreasing as K increases. On the other hand, we have shown that ρ^* increases in K in this case, which implies that $\frac{2-\rho^*}{1-\rho^*}$ must be increasing in K in this case. We can conclude from (3.55) that W_q^* must be decreasing as K increases in this case.

Now suppose that $\frac{1}{2(2-\gamma)} \leq s^* \leq \frac{1}{2-\gamma}$. In this case, the left side of (3.54) decreases in s. Since s^* increases in K, we can conclude from (3.54) that ρ^* must be decreasing in K in this case. The first-order condition (3.53) implies that

$$\bar{\lambda^*} s^* d[1 - (2-\gamma)s^*] = \frac{c(\lambda^* d)^2(2k^*\mu - \lambda^* d)}{k^*\mu(k^*\mu - \lambda^* d)^2} = c\frac{\mu^2}{d^2}(k^* W_q^*)^2(2-\rho^*). \tag{3.56}$$

As K increases, the left side of (3.56) decreases since s^* increases in K. On the right side of (3.56), both $(2-\rho^*)$ and k^* increase with K as ρ^* increases in K in this case. Thus, we can conclude from (3.56) that W_q^* must also be decreasing in K in this case.

Furthermore, we can show that

$$\pi^* = \lambda^* d(p^* - w^*) = \lambda^* d(\frac{1-\gamma}{2-\gamma} - \frac{c}{d} W_q^*). \tag{3.57}$$

Since λ^* increases in K and W_q^* decreases in K, we can conclude that π^* increases in K. Also,

$$p^* - w^* = (\frac{1}{\alpha^*} - 1)w^* = \frac{1-\gamma}{2-\gamma} - \frac{c}{d} W_q^* \tag{3.58}$$

Since w^* and W_q^* decrease in K, α^* is also decreasing in K.

(b) We first consider the case where s is fixed and show that the optimal k^* is increasing in γ.

Let k_0 denotes the optimal value of k when $\gamma = \gamma_0 \geq 0$. Then,

$$\Pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K} \right\} + \gamma_0 \frac{k_0}{K} G'^{-1}(\frac{k_0}{K}) = 0.$$

For any $\gamma_1 > \gamma_0$, we must have $\gamma_1 \frac{k_0}{K} G'^{-1}(\frac{k_0}{K}) > \gamma_0 \frac{k_0}{K} G'^{-1}(\frac{k_0}{K})$. Therefore,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K} \right\} + \gamma_1 \frac{k_0}{K} G'^{-1}(\frac{k_0}{K}) > 0.$$

Since $\Pi(k)$ is concave in k, the optimal k^* must be greater than k_0 when $\gamma = \gamma_1 > \gamma_0$.

Now consider the joint optimization problem of (s, k). Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $\gamma = \gamma_0$ and $\gamma = \gamma_1$, respectively. Suppose that $\gamma_1 > \gamma_0$. The two first-order conditions are given by

$$\frac{\partial \pi}{\partial k}|_{s=s_0,k=k_0} = -c\lambda d^2 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{\lambda=s_0\bar{\lambda},k=k_0} - \left[G^{-1}(\frac{k_0}{K_0}) + (1-\gamma_0) G'^{-1}(\frac{k_0}{K_0})\frac{k_0}{K_0} \right] = 0$$

$$\frac{\partial \pi}{\partial s}|_{s=s_0,k=k_0} = \bar{\lambda} \left\{ d \left[F^{-1}(1-s_0) - (1-\gamma_0) s_0 F'^{-1}(1-s_0) \right] - cW_q(s_0\bar{\lambda}, k_0, \mu, d) \right.$$
$$\left. -cs_0\bar{\lambda} \frac{\partial W_q(\lambda, k, \mu, d)}{\partial \lambda}|_{\lambda=\bar{\lambda}s_0,k=k_0} \right\} = 0$$

Let $k^*(\gamma, s)$ be the optimal value of k with fixed values of γ and s. As we have shown that the optimal k^* is increasing in γ for fixed s, we have $k^*(\gamma_1, s_0) \geq k^*(\gamma_0, s_0)$. As both $\frac{\partial W_q(s\bar{\lambda}, k, \mu, d)}{\partial \lambda}$ and $W_q(s_0\bar{\lambda}, k, \mu, d)$ decrease in k, we have

$$\begin{aligned} d\left[F^{-1}(1-s_0) - (1-\gamma_1)s_0 F'^{-1}(1-s_0)\right] - cW_q(s_0\bar{\lambda}, k^*(\gamma_1, s_0), \mu, d) \\ - cs_0\bar{\lambda}\frac{\partial W_q(\lambda, k, \mu, d)}{\partial \lambda}|_{\lambda=\bar{\lambda}s_0, k=k^*(\gamma_1, s_0)} \geq 0. \end{aligned}$$

Therefore, the optimal value of s must be greater than s_0 when $\gamma = \gamma_1$, i.e., $s_1 \geq s_0$. Also, $k_1 = k^*(\gamma_1, s_1) \geq k^*(\gamma_1, s_0) \geq k^*(\gamma_0, s_0) = k_0$. Therefore, both s^* and k^* are increasing in γ.

We can use (3.52) and (3.53) to cancel out γ and obtain

$$k = c\left(1 + \frac{\lambda dsK}{k^2}\right)\frac{\lambda d(2k\mu - \lambda d)}{\mu(k\mu - \lambda d)^2} = c\left(1 + \rho^2\frac{K\mu^2}{\bar{\lambda}d}\right)\frac{\rho(2-\rho)}{\mu(1-\rho)^2}. \tag{3.59}$$

Since k^* is increasing in γ and the right side of (3.59) is increasing in ρ, we can conclude that ρ^* is increasing in γ. Also, we can rewrite (3.59) as

$$c\left(1 + \rho^2\frac{K\mu^2}{\bar{\lambda}d}\right)\frac{2-\rho}{1-\rho}\frac{W_q}{d} = 1.$$

Since ρ^* is increasing in γ, we can conclude that W_q^* is decreasing in γ.

Finally, we can use (3.10) and (3.52) to obtain

$$c\frac{(2-\rho)}{\mu^2(1-\rho)^2} = (2-\gamma)Kw^2. \tag{3.60}$$

Since ρ^* is increasing in γ, the left side of (3.60) is increasing in γ. Since $(2-\gamma)$ is decreasing in γ, we can conclude that w^* is increasing in γ. □