INSEAD

The Business School for the World®

Incentive Design of On-Demand Marketplaces for Service and Innovation

A dissertation presented by

Konstantinos I. Stouras

to INSEAD faculty

in partial fulfillment of the requirements for the degree of

PhD in Management

March 2017

Dissertation Committee:

Serguei Netessine (chairman)

Karan Girotra

Andre Calmon

Steve Chick

This page is intentionally left blank

Dedicated to all my mentors
who passed me their passion for research
and who supported my work for the "next step"...

This page is intentionally left blank

Abstract

The rise of crowdsourcing marketplaces has allowed firms to involve large communities of external users (agents) in their internal processes. The aim of this thesis is the analysis of incentives and steady state dynamics of on-demand marketplaces in order to align the behavior of self-interested agents with the objectives of the marketplace.

In Essay 1, we examine the process of crowdsourcing innovation to an online community of solvers. Through a game-theoretic model, we examine the relationship between a seeker's choice of budget allocation across multiple awards and solvers' incentives for participation and effort. We characterize completely solvers' endogenous participation and (unobservable) effort decisions. The solvers compete only with those solvers who endogenously choose to participate, who are unknown to them ex ante participation. We show that multiple awards are required for sufficient solver participation. Finally, we prove that the seeker should optimally allocate all of her budget to the top participant even if she values multiple solutions from a large but finite population of solvers.

In Essay 2, we focus on the work-from-home (or virtual) contact center, a new type of the contact center business model. In a virtual contact center, demand is crowdsourced to a pool of freelance agents on priority, depending on their skill level and whether they have chosen to be available. Agent participation is voluntary and their idle time is not compensated. We study which priority classes partition generates the best incentives for agents of high ability to participate in order to maximize the profits of the firm. We show that discarding available information, or deploying a coarse priority scheme with two agent priority classes, maximizes firm profits and asymptotically maximizes social welfare. This provides a game-theoretical argument for the extensive use of coarse priorities by large-scale work-from-home service providers in practice.

In Essay 3, we study the online product support forum, an innovative business model for service. In a product support forum, the customer service of a firm is partially delegated to an active online community of users (firm's customers). We demonstrate that it may be to firm's best interest to strategically reduce its service capacity and to increasingly rely on its online community to serve its impatient customers on-demand. Our results shed light on why similar firms such as Microsoft and Apple, whose customers experience similar service needs, employ a fundamentally different strategy on the degree of their engagement in their respective online product support forums.

Keywords and phrases: on-demand marketplaces, service systems with random servers, server priorities, contests with uncertain number of competitors, incentives, all-pay auctions theory, innovation contests theory, self-confirming equilibrium, work-from-home contact centers, product support forums

This page is intentionally left blank

Research Output

Some ideas and figures have appeared previously in the following works:

K. I. Stouras. Product support forums: Customers as partners in the service delivery. *Available at SSRN 2868382*, 2016.

K. I. Stouras, K. Girotra, and S. Netessine. LiveOps: The Contact Centre Reinvented. *INSEAD Case Study*, 2014.

K. I. Stouras, J. Hutchison-Krupat, and R. O. Chao. Motivating Participation and Effort in Innovation Contests. *Available at SSRN 2924224*, 2017.

K. I. Stouras, S. Netessine, and K. Girotra. First Ranked First To Serve: Strategic agents in a service contest. *Available at SSRN 2696868*, 2016.

This page is intentionally left blank

Contents

List of Figures

This page is intentionally left blank

Introduction

> *We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover up all the tracks, to not worry about the blind alleys or describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work.*
>
> Richard P. Feynman (1918-1988), Nobel Lecture, 1966

Operations of modern marketplaces increasingly rely on incentives to better serve the incoming demand by the available supply. Precipitated by the lure of compensating for crowdsourcing service on-demand while gaining from differences in skill specializations and availability preferences, incentives for service expose a marketplace to numerous innate challenges. These challenges include misaligned objectives between freelance agents and the marketplace, vulnerability to provide stable service of high quality due to the fact that the participation decision of ability-heterogeneous agents is voluntary, and competition of a flexible labor capacity with the pre-determined staffing level of the marketplace. This dissertation examines the role of different operational levers in the efficient management of these challenges in attaining improved operational performance for on-demand marketplaces.

The process of sourcing a task to a crowd dates back to ancient times. Due to the account given by Herodotus (484-410 BC), the Father of History, it is believed that the ancient Babylonians had no doctors at all and instead they were crowdsourcing health care to their community. As Herodotus describes[1]:

> "They bring out all their sick into the streets, for they have no regular doctors. People that come along offer the sick man advice, either from what they personally have found to cure such a complaint, or what they have known someone else to be cured by. No one is allowed to pass by a sick person without asking him what ails him."

We would never know how many lives were saved by crowdsourcing health care in this way.

Another early example of crowdsourcing is attributed to Sir Francis Galton (1907), renowned anthropologist, biometrician and statistician. In 1906, Sir Galton visited an ox-weight-judging competition at the West of England Fat Stock and Poultry Exhibition in Plymouth. As Surowiecki (2005) describes:

> "A fat ox having been selected, competitors bought stamped and numbered cards, for 6*d.* each, on which to inscribe their respective names, addresses, and estimates

[1] Herodotus, *The Histories*, Book I (Clio), 440 BC.

> of what the ox would weigh after it had been slaughtered and "dressed". Those who guessed most successfully received a prize."

This idea that the crowd's average opinion often outperforms the opinion of any individual is another famous demonstration of the wisdom of the crowds (Surowiecki, 2005; Lichtendahl Jr et al., 2013).

With the advent of the Internet and the mass communication, crowdsourcing has found business applications ranging from generating innovation to connecting freelance agents with customers in order to provide service on-demand. Modern work-from-home contact centers allow independent contractors to flexibly choose when to work depending on their preferences and thriving online communities generate output that provides valuable service to other members of the community. The popular ride-sharing marketplaces of Uber and Lyft are yet another form of crowdsourcing service to an on-demand pool of agents competing for work.

To incentivize the agents, a marketplace monitors the output of the agents and shares with them their relative performance compared to their peers. In addition, the best performers are rewarded in the form of monetary or non-monetary benefits.

Due to their similarity with sports competitions, patent races or rent-seeking competitions where contestants are competing for some rewards, settings in which agents compete for innovation or service are naturally termed *contests*. In fact, websites that elicit innovation from the crowd such as InnoCentive.com and 99designs.com label them as such. As a concrete example, consider the case of InnoCentive.com, in which a firm is seeking a solution to a problem by exposing it to an active online community of crowd-solvers. The solvers submit solutions to the posted problem and the best of them receive some monetary rewards and accumulate reputation points that lifts their profile in the social ladder of the community.

Online crowdsourcing contests differ from traditional contests in the following ways. First, the pool size of potential contestants in online crowdsourcing contests is large, and lies in the order of ten thousands. For instance, the work-from-home contact center LiveOps outsources demand to an army of twenty thousand independent contractors who compete for calls (LiveOps, 2014). Second, whereas a traditional contest organizer can control the number of contestants to enter the contest, agent participation in online service marketplaces is not guaranteed by the very nature of the contract between the marketplace and the agent. For example, drivers of the popular ride-sharing platform Uber are free to determine their own work schedule and cannot be called Uber employees in legal terms. Third, the agents of a marketplace exhibit a significant heterogeneity in terms of their skills that make them more (or less) capable to handle a specific service request. Finally, if and when they decide to work, the agents must be induced to work hard (i.e. exert effort) to deliver the output needed by the marketplace.

This dissertation focuses on the best way to crowdsource demand to a pool of freelance agents who create value to the marketplace on-demand, when agent participation is voluntary.

Contests and all-pay auctions

A contest can be represented as an auction in which the higher "bidder" (i.e. the contestant who exerts the highest effort) receives the most valuable "object" (i.e. the award with the highest

value) and similarly in case of multiple awards, but all bidders pay their bid to the "auctioneer" (i.e. the contest organizer). A formal connection between contests and all-pay auctions has been established by Siegel (2009) and has found applications in crowdsourcing, R&D races and rent seeking (see Konrad (2009) and Dechenaux et al. (2015) for literature reviews).

We build on the incomplete information, all-pay contest models of Moldovanu and Sela (2001) and Moldovanu et al. (2007) accounting for the *voluntary participation choice* of the competing contestants. In particular, Moldovanu and Sela (2001) consider a contest in which agents' participation is guaranteed and ability-heterogeneous contestants are competing by bidding costly effort. Moldovanu and Sela (2001) show that a resource-constrained contest organizer should optimally allocate its entire budget to the highest effort agent in order to maximize the expected total effort of the agents. This is often referred to as the *winner-takes-all* (WTA) allocation. In a similar model, Moldovanu et al. (2007) find that when agents are instead competing for status, the most hierarchical formation maximizes the total expected effort, when agent participation is exogenously fixed and the ability distribution is sufficiently convex.

Modeling sales agents competing for sales in a variety of demand territories, Kalra and Shi (2001) study the optimal *sales contest*. Kalra and Shi (2001) consider ability-homogeneous agents who exert sales effort and the amount of sales they generate is affected by random demand shocks. Similar to Moldovanu and Sela (2001) the authors find that WTA is optimal, i.e. allocating the entire budget to the agent with the highest sales maximizes the total expected sales generated.

Terwiesch and Xu (2008) and Körpeoğlu and Cho (2017) study innovation contests with ability and effort, however, they do so normalizing agent participation cost to zero, leading to a WTA award scheme. The seemingly innocuous assumption regarding agent participation cost fails to address agents' voluntary participation choice, which is central in online crowdsourcing marketplaces. As a consequence, all agents enter the contest and the innovation contest organizer does not face a trade-off between how many solvers and of what skill level to encourage to enter, and how to motivate those solvers that do enter to work hard.

Overall, prior literature that has shown the optimality of the WTA allocation, it has identified four reasons for the existence of multiple awards in practice: sufficiently convex cost of effort (Moldovanu and Sela, 2001), sufficiently convex ability distribution (Moldovanu et al., 2007), risk aversion (Kalra and Shi, 2001), or the need to induce search (Erat and Krishnan, 2012). We note that all these papers provide also sufficient conditions for the optimality of a WTA allocation.

Nevertheless, recent empirical findings of Kireyev (2016) suggest that multiple awards exist in practice while there are no significant indicators of risk aversion among the contestants and assuming linear or even quadratic costs of effort. Kireyev (2016) attributes his empirical findings to the existence of information asymmetry between the agents and the contest organizer. In particular, the author shows that under complete information a winner-takes-all allocation is optimal, but when there is information asymmetry multiple awards are beneficial.

We complement the literature on contests and we show that the fact that agent participation is voluntary may drive the prevalence of multiple awards in online crowdsourcing marketplaces. We propose a novel theoretical framework to model the trade-off between participation and

effort of the agents. We show that such a trade-off forces the contest organizer to provide multiple awards.

Applications of contest theory in operations management

We study three modern business models under the lens of operations management (Girotra and Netessine, 2014) that challenged the traditional way to generate innovation and provide service for a firm leveraging large communities of external users (von Hippel, 2005, 2016):

1. *Crowdsourcing innovation.* Traditionally, the innovation process was conducted entirely in-house. However, the rise of crowdsourcing has allowed firms to increasingly involve large communities of external users in their internal innovation processes (von Hippel, 2005; Terwiesch and Ulrich, 2009; von Hippel, 2016). This posed new challenges for the firms to motivate high engagement by the users and incentivize them to work hard, so that the firm benefits from crowdsourcing innovation to a large heterogeneous pool of outside users. The effective management of an innovation contest is of significant managerial value.

2. *Work-from-home contact centers.* Call centers have been used extensively to provide service to customers of an organization (Gans et al., 2003). The advent of the internet created a plethora of channels that customers choose to interact with the focal firm including video call, real-time messaging capabilities, social media and e-mail among others (Stouras et al., 2014). To cope with this multi-channel transformation of the service landscape, modern service providers employ a work-from-home business model of the contact center as a cost-effective alternative.

3. *Online product support forums.* Recently, organizations ranging from newly created startups to Fortune 100 corporations such as Microsoft, Apple and PayPal increasingly crowdsource service to their active online community of experts via an online product support forum. Such an online community is composed by customers and members of the overall ecosystem of the respective service provider, so the firms are essentially crowdsourcing service back to their customers! Online product support forums have the potential to provide fast and reliable service leveraging the expertise of the community and even resulting in second-order branding benefits for the focal firm.

This dissertation takes the first step in understanding the incentive design of these marketplaces proposing a novel theoretical framework to model the voluntary participation choice of the agents who provide service in all of the above marketplaces. Chapter 1 of the dissertation studies the Business Model Innovation 1 by building the first innovation contest model in which agent participation is endogenous. This model generalizes existing contest theory that considers an exogenously fixed number of participants, whereas it provides a theoretical foundation that does not contradict with recent empirical findings. While Chapter 1 focuses on crowdsourcing innovation, Chapter 2 and Chapter 3 are centered around crowdsourcing service. Chapter 2 focuses on the Business Model Innovation 2 and characterizes the optimal way to allocate demand to participating agents on priority to maximize firm profits. Our characterization

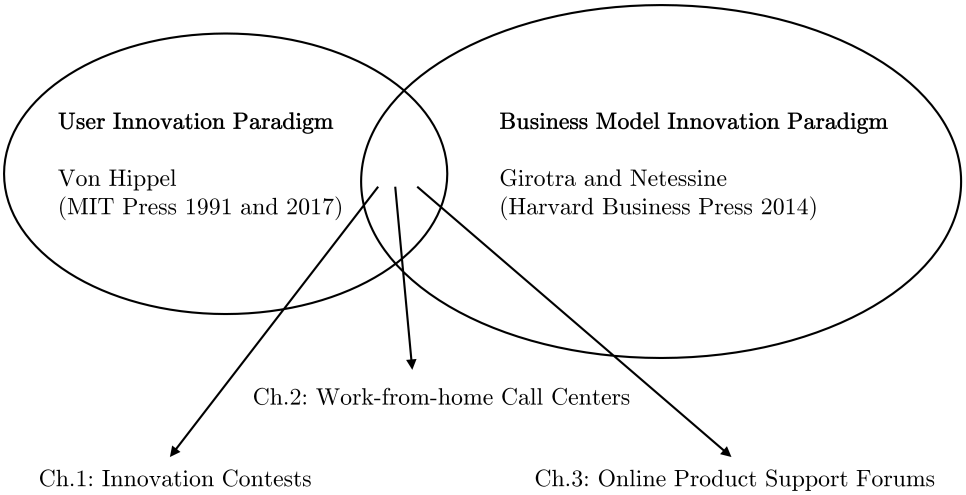


Figure 1 Thesis overview: Work is carried out in the "household sector" of national economies, by the users of large online communities who generate output voluntarily.

provides a causal explanation for why work-from-home contact centers prefer to rank their agents in a few coarse categories by their sales output. Chapter 3 studies the Business Model Innovation 3 and explores the strategic behavior of the users of the online community of a firm which service is partially delegated to. The third chapter also illustrates the usefulness of modeling the voluntary participation choice of the users by providing critical strategic recommendations to firms which employ a product support forum for service.

Figure 1 provides a high level overview of the contribution of this thesis which lies in the intersection of the Business Model Innovation Paradigm (Girotra and Netessine, 2014) and the User Innovation Paradigm (von Hippel, 1991, 2017) under the lens of Operations Management. A brief account of the individual chapters is provided below.

Innovation contests

Innovation contests are competitive settings in which an innovative solution to an existing problem of an organization is crowdsourced to a pool of external users. A resource-constrained firm (seeker) decides on the best way to use its available budget by offering a distribution of rewards to the best performing solutions. Freelance agents (solvers) observe the promised reward distribution and determine whether or not to participate and how much effort to exert to enhance their solution, conditional on participation.

Purely effort-driven contests have been studied extensively in Economics (see the literature reviews of Konrad (2009) and Dechenaux et al. (2015)) and recently have been applied in Operations Management (Terwiesch and Xu, 2008; Terwiesch and Ulrich, 2009; Ales et al., 2017; Nittala and Krishnan, 2016). A notable exception is the recent work of Körpeoğlu and Cho (2017) that corrects mistakes in the equilibrium analysis of ability and effort model of Terwiesch and Xu (2008). However, both Körpeoğlu and Cho (2017) and Terwiesch and Xu (2008) fail to address the voluntary participation decision of the agents in an innovation contest when solving firm's problem.

Pre-announced contest characteristics	Min	Median	Mean	Max
Number of prizes	1	4	5	50
Total budget of seeker	\$500	\$1,000	\$1,450	\$10,000
Participating solvers	58	187	193	499
Solver population per contest	58	235	247	623
Participating solvers/ Solver population per contest	37.23%	78.48%	77.53%	100%

Figure 2 Summary of contests' characteristics of Tongal.com during 2011-2015 as reported by Kireyev (2016).

In Chapter 1 we propose a novel theoretical framework to analyze the equilibrium behavior of strategic agents utilizing the powerful notion of self-confirming equilibrium (SCE) of Fudenberg and Levine (1993a). We note that the SCE is a generalization of Nash Equilibrium and it has been shown to coincide with the convergence of a learning dynamics process to a steady state (Fudenberg and Levine, 1993b). We model this steady state of the system at which the agents make a rational participation and effort decision based on beliefs on the anticipated actions of their peers. Such beliefs are distributions on the actions of the others conditioned on an agent's own actions. In a SCE we require agent's *ex ante* beliefs to not contradict with the observed *ex post* equilibrium outcomes of the innovation contest. Further, we explicitly model the number of participating agents as a non-negative random variable. We show that in a SCE with symmetric strategies the number of participating agents follow a Binomial distribution with a participation probability that is determined by the endogenous participation choices of the strategic agents.

Having characterized the equilibrium behavior of the agents, we next demonstrate that multiple awards can be optimal when agent participation is voluntary. Indeed, to motivate sufficiently high participation and effort, marketplaces that run innovation contests such as Tongal.com often announce *multiple* awards as many as 50 awards with a median of four awards (see Table 2 of Kireyev (2016) reproduced in Figure 2, and see also §D.2 on p.102 for data collected from InnoCentive.com and Kaggle.com). As Kireyev (2016) reported to us by personal communication solver participation in contests in Tongal.com is not guaranteed and the fraction of participants fluctuates with a median of 78% of the entire population. Our model continues to be valid for an innovation contest with exogenously guaranteed agent participation, and it simplifies to the existing works of Moldovanu and Sela (2001) and similar literature in this vein. In addition, our results do not contradict the empirical findings of Kireyev (2016) and Boudreau et al. (2011), and the experimental works mentioned in Dechenaux et al. (2015) that show that a variety of contests in practice use multiple awards (as opposed to a winner-takes-all design).

Chapter 1 further makes a conceptual contribution to the contest literature. It provides new insights on the way the agents substitute costly effort for their intrinsic ability. In particular, Chapter 1 shows that there is a group of low ability agents for whom ability and effort are *complements*, that is the higher the ability of an agent, the higher effort he exerts in equilibrium. Interestingly, we show that there is an agent (the "hardest worker") so that agents of higher

ability than him *substitute* effort for their ability to the extent that the most capable agent exhibits a decrease in his effort exerted. This novel effect is shown in the conservative case of linear cost of effort and is further exacerbated for a convex cost of effort. We note that agents' performance, which is a convex combination of ability and effort, is always strictly increasing in ability in equilibrium.

In Chapter 1 we develop a new theoretical foundation of contest theory to incorporate agents voluntary participation decision and costly effort choice. Then, we apply this framework to two novel crowdsourcing service settings.

Work-from-home contact centers

Chapter 2 concerns a work-from-home service marketplace which ranks its pool of freelance agents in a predetermined number of priority classes based on their sales performance (see Business Model Innovation 2 above). The agents are paid only for the amount of time they are utilized and higher sale performers earn weakly more. The objective of the marketplace is to maximize profits by allocating work to participating agents on-demand. Due to the similarity with a contest, we term such a setting a *service contest* and apply the techniques of Chapter 1.

A work-from-home contact center can be conceptualized as a service marketplace in which independent contractors are competing to serve available demand, and agent participation is voluntary. In particular, the agents are ranked by ability which is a proxy for the amount of sales they can generate, and they are free to form their own work schedule by choosing among available work shifts in advance. What is different compared to a regular contest is that in a service contest the "rewards" (i.e. available routed demand) promised to the agents are *stochastic*. Indeed, the firm does not compensate the agents for their time staying idle in the system and hence the earnings of the agents depend on the incoming traffic that is only shared among any participants.

In order to make a rational participation decision, each work-from-home agent faces two kinds of uncertainties: he is unsure of the number of the agents who will choose to participate, as well as of his own rank-order among the participants. To model this "strategic uncertainty", we follow the techniques of Chapter 1 and assume that the agents form (ex ante) beliefs about the anticipated participation actions of the others. In a SCE we require the beliefs of the agents to not contradict the observed outcome. That is, the beliefs are self-confirmed in equilibrium.

We next describe some challenges we initially encountered with modeling this setting and contrast our contributions with the existing literature in Operations Management.

Concurrently to our work, Gurvich et al. (2015) study the optimal pricing scheme to alleviate the possible externalities the agents impose to their peers when too few or too many participate in the system. The model of Gurvich et al. (2015) is similar to ours: there is a population of N agents and incoming demand λ. At the beginning of the period the agents decide to participate and cannot exit the system during a sufficiently long period of work (so that a long-run behavior to have some meaning). Gurvich et al. (2015) further extend their model to a multi-period setting.

In our setting, we explicitly model the number of participating agents as a non-negative

random variable $\mathcal{N}$ and show that in a SCE with symmetric strategies it follows the Binomial distribution with a probability parameter being a decision choice of the agents. Gurvich et al. (2015) approximate the expected amount of time an agent is busy, a notion we refer to as the *expected utilization* of the agents, with the ratio of the expected demand over expected number of agents (or $\frac{\lambda}{\mathbb{E}[\mathcal{N}]}$). Unfortunately, we show that the approximation $\frac{\lambda}{\mathbb{E}[\mathcal{N}]}$ of Gurvich et al. (2015) of the *actual* expected utilization $\mathbb{E}\left[\frac{\lambda}{\mathcal{N}}\right]$ cannot be applied to our setting for the following two reasons. First, using Jensen inequality the latter is greater or equal to the former. Hence, a condition $\frac{\lambda}{\mathbb{E}[\mathcal{N}]} < 1$ does *not* guarantee a stable system. We note that the forthcoming paper of Cachon et al. (2017) also follows Gurvich et al. (2015) and makes the same approximation.

Second, although one can show that $\mathbb{E}\left[\frac{\lambda}{\mathcal{N}}\right]$ indeed converges to $\frac{\lambda}{\mathbb{E}[\mathcal{N}]}$ as $N \to \infty$ under some assumptions, these assumptions do not model the voluntary participation dynamics of agents in a two-sided marketplace. We summarize sufficient conditions for the latter expressions to coincide as agent population size grows without bound in the following proposition[2].

Proposition 1. *Assume that:*

(1) First, Poisson demand arrives with rate λ and then $\mathcal{N}$ agents participate all at once, at the beginning of a single period with infinite duration. Once an agent enters the system, he cannot exit.

(2) $\mathcal{N} \sim$ Binomial(N, p)

(3) All agents participate with the same fixed probability $p > 0$, which does not depend on N.

(4) There is always one additional agent in the system (inflexible employee).

Then, we have that

$$\begin{aligned} \lim_{N\to\infty} \mathbb{E}\left[\frac{\lambda}{\mathcal{N}+1}\right] &= \lim_{N\to\infty} \sum_{k=1}^{N} \frac{\lambda}{k} \left(\begin{array}{c} N \\ p \end{array} \right) p^k \, (1-p)^{N-k} \\ &= \lim_{N\to\infty} \frac{\lambda}{\mathbb{E}\left[\mathcal{N}\right] + p} = \lim_{N\to\infty} \frac{\lambda}{p\,(N+1)} = 0 \end{aligned} \tag{0.1}$$

Proof of Proposition 1. The proof follows by Cribari-Neto et al. (2000) and the SLLN. □

The applicability of Proposition 1 is very limited for the following reasons. Assumption (1) models the voluntary participation choice of the agents but fails to account for the *voluntary exit* of the agents from the system. Although we show that the Binomial distribution indeed arises in a symmetric self-confirming equilibrium giving support to assumption (2), the assumption (3) does not address the intuitive fact that the participation probability *should* depend on the population size. In practice, the probability to enter given that ten agents could participate should be larger than the probability to enter the system when one thousand agents could enter the system. In fact, we show in Stouras et al. (2016) that agents' participation probability decreases at the order of $O\left(\frac{1}{N}\right)$. What is more, assumption (4) is necessary for (0.1) to hold. Technically, this implies that the support of the (random) number of participating agents does not include the zero *by assumption*. Without assumption (4) we have that $\mathbb{E}\left[\frac{\lambda}{\mathcal{N}}\middle|\, \mathcal{N} \geq 0\right] = +\infty > \frac{\lambda}{\mathbb{E}[\mathcal{N}|\mathcal{N}\geq 0]} = \frac{\lambda}{Np}$ for all $N \geq 1$. However, the popular ride-sharing marketplaces of Uber

[2]The author thanks Steve Chick, Rouba Ibrahim, Ioannis Panageas, John Tsitsiklis and Amy Ward for fruitful discussions on stability.

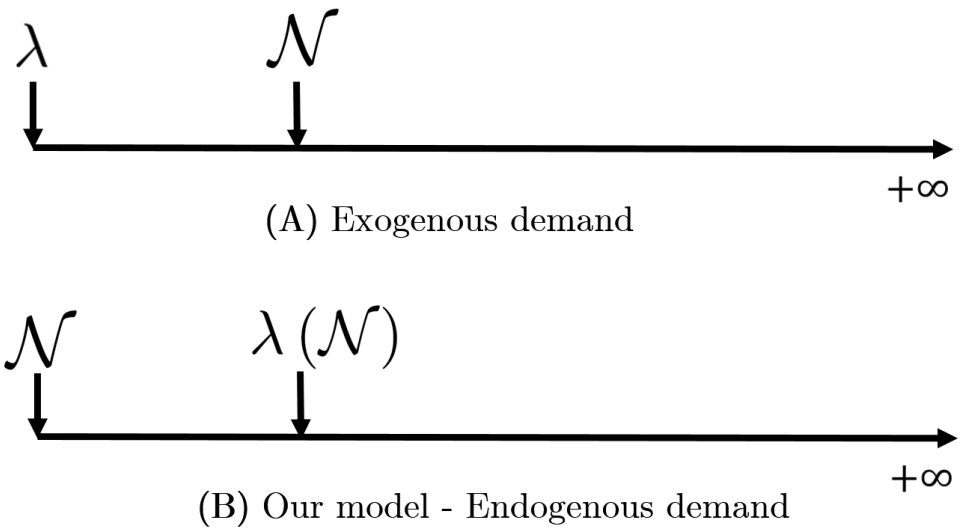


Figure 3 (A) A model of an on-demand service platform with exogenously fixed demand. (B) Our model: There is a single period of work of infinite duration in which first the agents decide to participate and stay in the system for the entire period, and then demand is realized.

and Lyft do *not* have any fixed workforce and operate by sourcing work to freelancers who self-determine whether to participate or not depending on their own schedule.

In addition, we note that any model with exogenously fixed demand that satisfies the assumption (1) of Proposition 1 is unstable with positive probability (see Figure 3(A) for an illustration). In particular, there is a strictly positive probability that the participating capacity is insufficient to exceed the average demand. This leads to an unstable system with positive probability $\mathbb{P}[\mathcal{N} \leq \lambda] > 0$. In the special case that the random capacity follows a Binomial distribution (i.e. it satisfies the assumption (2) of Proposition 1) one can use Chernoff bounds and show that the probability $\mathbb{P}[\mathcal{N} \leq \lambda]$ is "small" and can be controlled by the firm. Thus, one could focus on an on-demand service platform that is (ex ante) stable with high probability. Unfortunately, the fact that there is a positive, albeit tiny, probability having an unstable system, implies that the expected waiting time that the customers face is infinite[3]!

We are aware of *three* ways to overcome the aforementioned challenge of stability. First, one can consider customer abandonments which is a natural assumption to make. As Baccelli et al. (1984) shows when there is a positive customer abandonment probability, the system is always stable. However, customer abandonments make the exact analysis of the steady state intractable and queueing scholars apply fluid approximations at an appropriate scale of the system. To our knowledge, only Ibrahim (2015) has succeed in formulating fluid approximations for Binomial servers so far. We note that Ibrahim (2015) assumes that the agents' participation probability is *exogenously fixed* and does not depend on the size of the population. Unfortunately, as we show this assumption does not apply to our setting.

Second, one can model the exit decision of the servers and scale the system so that the incoming rate of arrivals of customers and rate of participating agents are in a steady state. This is indeed the approach followed by Banerjee et al. (2015) leading to a stable system. However,

[3]To see that, calculate the expected waiting time by conditioning on the system being stable.

work-from-home contact centers restrict any agents who choose to work at a predetermined work shift to stay in the system for the entire work shift. In practice, such a work shift typically lasts six to eight hours Stouras et al. (2014). In addition, in our setting there are *agent priorities* which render the exact analysis intractable following this modeling approach.

Third, one can endogenize the demand arrival process so that it leads to a stable system in equilibrium (see Figure 3(B) for an illustration). Taylor (2016) first applied this modeling assumption in two-sided marketplace and studied its optimal dynamic pricing strategy, while Tang et al. (2016) adopt similar techniques on a ride-sharing platform. We follow Taylor (2016) but we explicitly model the number of participating agents as a *non-negative random variable* and show that in a symmetric SCE it follows the Binomial distribution with a probability parameter which is an endogenous choice of the agents. We extend Taylor (2016) using *expected utilizations* and we analyze the steady state dynamics of a markovian queueing system with random servers and any number of server priority classes. While Taylor (2016) and Tang et al. (2016) consider homogeneous agents with heterogeneous participation cost, we model ability-differentiated agents who share the same fixed cost to work for a give work shift. Our analysis can be extended to the case where the outside option (i.e. opportunity cost) of the agents is strictly increasing in their ability such that the ratio ability over opportunity cost per agent is strictly increasing in ability.

The self-confirming equilibrium theoretical framework of Chapter 1 directly applies to a work-from-home contact center setting to model the endogenous participation decision of the work-from-home agents of Chapter 2. Next, we study a similar crowdsourcing system focusing on the dynamics of agents' participation decision to provide service.

Online product support forums

The idea behind Chapter 3 leading to the solo-authored work Stouras (2016) came when the author of this dissertation was teaching the fundamentals of the Core Operations Management course at the INSEAD MBA Programme in March-April 2016. In particular, the author was responsible for more than 300 INSEAD MBA Students who were frequently sending emails related to the course logistics, the cases, the exams, and even shared interesting business model innovation examples from their practical experience. The author utilized Yammer, an online social networking platform which can be used internally only by INSEAD members[4]. The MBA Students were asked to subscribe to notifications from this group and to regularly visit it for announcements and useful material related to the OM course. The author cultivated an online community on Yammer encouraging the Students to post questions and motivating them to *respond* to questions posted by their peers. Given that it is almost impossible to handle spikes in email demands from the large community of 300 MBA Students, this crowdsourced solution provided remarkably fast and reliable service to the class on-demand and with superior quality. Essentially, the same business model innovation for service support is currently in use by large corporations such as Microsoft, Apple, PayPal and Walmart, as well as various start-ups which

[4]See the Yammer POM 16D group available at:
https://www.yammer.com/thelearningnetwork/#/threads/inGroup?type=in_group&feedId=7492861.

employ an online product support forum instead of a traditional contact center for their service support.

Chapter 3 concerns a firm which partially delegates customer service to an active online community of users (see Business Model Innovation 3 above). Impatient askers (i.e. customers) post easy and hard questions and receive service in the form of answers either by an online community of users (i.e. other customers), or by the servers of the firm. The available staffing level in the presence of an online community is a choice of the service provider, whereas user participation is voluntary and costly, and users benefit from providing service on-demand in the form of reputation points accumulated over time.

Given a rate that firm's servers respond to questions as a Stackelberg leader, the users of the online community follow by choosing their service rate. We explicitly account for the non-participation option of the servers and the users and we allow the service rate to be zero in this case. We show that a sufficiently active server can discourage users from participating into the online product support forum. Interestingly, we show that the users service rate for both easy and hard questions is unimodal in firm's service rate. That is, there is an initial range of service rates of the firm that the service rates of the two competing parties behave as complements and the users respond faster in response to a firm with a faster service rate. However, we demonstrate that after a "tipping point" these quantities become substitutes at a different level for each question type.

Further, we show that despite any available high-cost-high-reward hard questions, the users mix their responses and often reply to low-cost-low-reward questions. We term such "mixed" equilibrium behavior as *exploration* to reflect the fact that the users respond to both types of questions with positive probability. For a sufficiently active firm the users' participation cost of resolving an easy question offsets any potential awards of reputation benefits for easy questions, and the users cluster their responses only under any high-cost-high-reward hard questions available. In that case we say that users perform *exploitation*, i.e. they only respond to the type of questions with the highest potential. An exploitation equilibrium outcome may be particularly inefficient when easy questions are swarming the system and outside users choose to resolve only the spare hard ones.

We characterize firm's optimal staffing level in managing an online product support forum. We solve firm's profit maximization problem as a function of askers' impatience level. Interestingly, we find that askers' value is not always increasing in firm's service rate. This implies that it may be to the best interest of the firm to strategically *reduce* its service rate to boost a faster response rate from the community. The latter insight offers a game-theoretic explanation for why companies such as Microsoft and Apple with similar products and large online communities manage their online product support forums differently. We have collected data from these respective online communities and we aim to test these hypotheses on empirical grounds.

By providing critical strategic recommendations to firms which employ a product support forum for service, Chapter 3 elucidates the usefulness of modeling the voluntary participation choice of the users of the online community. The framework developed in Chapter 3 can also be applied to other settings in which contestants choose to enter a contest among many contests that run in parallel and have a different deadline. Chapter 3 also complements Chapter 1 and

Chapter 2, by dealing with the dynamic (as opposed to one-shot) participation decision of the agents who co-exist with available staffing level of the focal firm (as opposed to an entirely crowdsourced workforce). In the subsequent sections we describe the individual chapters in detail. Note that all the three chapters are self contained and can be read in any order.

Organization of the Thesis

We separate the three Chapters of this Thesis into two Parts: Crowdsourcing Marketplaces and Service Marketplaces. Chapter 1 of Part I examines a traditional innovation contest setting accounting for solver voluntary participation decision. Chapter 2 and Chapter 3 of Part II analyze a model of the work-from-home contact center and the online product support forum, respectively. All proofs are relegated to the Appendices. In Appendix A we present a case study of LiveOps, a start-up that motivated the research of Chapter 2. A table of notation used in Chapter 1 and Chapter 2 is contained in Appendix D and Appendix E respectively. Appendix C compares two popular contest models and shows that they lead to qualitatively similar equilibrium actions of the agents. Finally, we conclude with a brief review of our contributions, directions for future research and implications of this dissertation for practitioners and the future of work.

Acknowledgments

I would like to express my gratitude to my mentors, Serguei Netessine and Karan Girotra, whose encouragement and generous support during my doctoral studies have been invaluable to me. My dive into the realm of on-demand marketplaces started on the first month of my doctoral studies, on a rainy morning in Fontainebleau on Friday, September 16, 2011. That day I discussed with Serguei two research ideas. The first idea was about the business model of a work-from-home contact center start-up called LiveOps and the second idea concerned GroupOn. I was immediately intrigued to learn everything about LiveOps. I was fascinated by the fact that LiveOps looked like a "service system upside down" in which the *servers* are queueing for work, instead of the customers as it is traditionally assumed by the queueing theory to date. Late that evening, I also discussed with Karan who shared my enthusiasm that working on LiveOps would be a promising project. And just like that, I started working on what I now feel is the "future of work".

I am very appreciative to Serguei and Karan for mentoring me through this journey, and for guiding me to conduct original research at the frontiers of our body of knowledge under the lens of Operations Management. The never-say-no and big picture thinking of Serguei, and the sharp intuition and honest critique of Karan constantly challenged me to improve my academic acumen. I feel thankful to Serguei for his strong support of my "crazy" idea to create *OperationsAcademia.org*, a marketplace for the Academic Job Market in Operations (and related fields and disciplines); many things could go wrong but this initiative has already started to bear fruits and attracts global attention. Being advised by Serguei and Karan, I have matured in so many dimensions including my approach and taste on research, how to present

and communicate my ideas, my teaching style, and my overall profile as an academic scholar in a business school. I will always be indebted to them.

I am also deeply grateful to Steve Chick, who has been always available and honest about my work. I was very fortunate to have been taught the fundamentals of Stochastic Processes by him. Steve combines the rigor and clarity of a mathematician with a practice-driven mindset. I learned a lot by his side during our interactions for the Beer Game Exec Ed courses he taught and the on-site visits to L'Oreal and Bosch in Rambouillet and Rodez, France respectively. I thank Steve for the helpful discussions on "stability" and for pushing me to focus on the "forest" rather than the "tree".

Moreover, I was very fortunate to have met Andre Calmon during his first year as a Faculty member at INSEAD. Andre is always approachable and has been a great co-advisor and even a friend at some tough times for me. I was honored to have been an Instructor for his POM Core MBA course at INSEAD and I learned a lot from his approach on research and teaching.

Special thanks to Jürgen Mihm for introducing me to the field of New Product Development and the area of Innovation Contests. Indeed, the idea of modeling LiveOps as a "service contest" matured through his PhD classes which were full of research ideas and one of the most thought-provoking classes I had at INSEAD. I am blessed Jürgen's office door was always open for me to discuss some conceptual research ideas or simply to chat about "how is life".

I will never forget the time Enver Yücesan spent with me discussing an earlier version of my work on LiveOps. Enver was so kind and helpful to me, and we had many helpful discussions in his office.

Also, I am honored to have been taught by Luk Van Wassenhove. I admire Luk a lot; he is a leading management thinker and one of the greatest researchers in Operations Management. I thank Luk for attending my mock job talk and sharing his wise insights on my work back in September 2015.

I also want to thank all my other colleagues at INSEAD, the Administration and the PhD Office and, above all, the members of the TOM Area for the unique, friendly and stimulating atmosphere they created. Special thanks to my teachers Yale T. Herer (Technion - Israel Institute of Technology), Sameer Hasija, Hao L. Lee (Stanford Graduate School of Business), Ehud Lehrer (Tel-Aviv University), Dana Popescu, Nils Rudi and Manuel Sosa for what I learned from their classes.

I am indebted to Raul Chao and Jeremy Hutchison-Krupat for their continuous guidance and support at the Darden School, University of Virginia. I also thank Casey Lichtendahl for fruitful discussions on tournaments and research on forecasting.

I thank John N. Tsitsiklis for fruitful discussions during my visits at MIT.

I am grateful to Soo-Haeng Cho for pointing out an error on an earlier submission of my work to the MSOM Student Competition in 2015. My LiveOps paper has been benefited by the discussions with Javad Nassiry, Rouba Ibrahim and Amy Ward who have all contributed to my understanding of the problem I study through multiple correspondence we exchanged.

My friend Ioannis Panageas deserves an acknowledgments section on his own. First and foremost, I thank him for being a true friend during the bad and the good times. Devastated by the slow progress of my work during the 4th year of my doctoral studies, Ioannis was there for

me to remind me to not give up, to stay focused, to push deeper and to work harder. We spent long Skype calls together chatting about our daily life and beautiful mathematical theorems such as Brouwer's *Invariance of Domain Theorem* and techniques from Dynamical Systems. My work has been improved significantly from Ioannis' sharp comments during my visit at Georgia Tech in 2016 and MIT in 2017 respectively. I thank Ioannis from the bottom of my heart for his support during the writing of this thesis.

Many thanks to the Greek gang: Andreas Chouliaras (University of Reading, Henley Business School), Dimitris Gkounis (ETH), Nikos Karianakis (UCLA), Ioannis Lionis (Apttus, San Francisco), Alexandros Mavridis (Abbott, Paris), Charis Petroxeilos (University of Athens), Ioannis Siskos (KLU, Hamburg), Thanos Tsamis (European Court of Auditors, Luxembourg) and George Tzallas-Regkas (Accenture, Dubai) for being there when I needed them and for all the great moments we shared over the last 5 years. I am blessed to have met Jacquiline Sclavou and I thank her for her selfless love. I had a great time sharing my accommodation and traveling together with Panos Markou (IE, Madrid) during each Conference of the last years.

Rolf Höfer made my first years at INSEAD unforgettable. No day was ever the same for me from the moment he moved to Singapore.

Special thanks to my mentor Themistocles M. Rassias for his support all these years and for all the things I learned from my interaction with him.

I thank Stelios Kavadias for all the intellectual discussions we had and for always remembering me when he was visiting INSEAD to teach for as shortly as just a weekend.

Additional thanks go to my PhD buddies Matthijs, Afonso, Ruslan, Kate, Jeeva, Ashish, Kashish, CK, Maciej, Sergio, Simona, Sorah, Gloria, ChunLiu, Karca and Simone for their support and the fun time we had together. I thank all the INSEAD MBA Students I met, in and out of classrooms, over the last years in France and Singapore. The interaction with them taught me so much and they broadened widely my horizons for culture and diversity.

Certainly, many others I met during my INSEAD doctoral studies are left out here but I hope they know I enjoyed my time with them and I am honored I met them.

For my father Ioannis and mother Chryssa I am truly grateful. They raised me with sincere love, passed me a strong will and dedication to what I do, and wholeheartedly supported my choices in life. I am proud of my sister Athena for how mature she is and for having found her "path" in life. Despite the moments of crisis Greece is now facing, I hope we will soon live closer and see each other often. A warm hug to my aunt Anastasia for her love and support.

Finally, I send my deepest gratitude and love to Linda for being so understanding and supportive over the last years. We certainly had to cut back a lot while living in different continents for long periods, but in the end I had the incredibly rare luck of having a person who loves me unconditionally and accepts my commitment for deep work.

Singapore, March 2017

Konstantinos I. Stouras

Part I.

Crowdsourcing Marketplaces

This page is intentionally left blank

Chapter 1

Motivating Participation and Effort in Innovation Contests[1]

This Chapter studies innovation contests, a business process through which a firm (seeker) crowdsources innovation to a large pool of users (solvers). Through a game-theoretic model, we examine the relationship between a seeker's choice of budget allocation across multiple awards and solvers' incentives for participation and effort. We investigate "contest specialization", a structural characteristic of each contest that defines the degree to which solvers can substitute ability for effort to enhance the performance of their solution. We characterize completely solvers' endogenous participation and effort decisions. The solvers compete only those solvers who choose to participate, who are unknown to them ex ante solvers' participation decision. We show that multiple awards are required for sufficient solver participation. Finally, we prove that the seeker should optimally allocate all of her budget to the top participant even if she values multiple solutions from a large but finite population of solvers.

Key words: open innovation; crowdsourcing; incentives; endogenous participation; innovation contests with uncertain number of solvers

[1]This Chapter is based on joint work with Jeremy Hutchison-Krupat and Raul O. Chao (Stouras et al., 2017).

1.1 Introduction

The rise of open innovation and crowdsourcing has allowed firms to involve large communities of external users in their innovation process (Terwiesch and Ulrich, 2009). Unfortunately, user (solver) participation in such innovation contests is not guaranteed. Solvers, who have different skill or ability levels, may find the cost of participation prohibitive. If and when they decide to participate in a contest, the solvers must be induced to exert effort that delivers the output needed by the firm (seeker). To incentivize the solvers, a resource-constrained seeker faces a trade-off between allocating fewer awards of larger value versus more awards of smaller value. Most papers in the existing literature focus on incentives for effort alone and they largely find that a winner-takes-all award scheme is optimal. In contrast, we establish that *multiple awards* are needed to balance solver incentives for participation and effort in settings where solver participation is voluntary.

In practice, crowdsourcing a task to outside solvers comes with two main challenges related to the incentive design of the award scheme. First, crowdsourcing contests can differ widely in structure depending on the degree to which ability and effort (together) determine output, an intrinsic characteristic of each contest we refer to as *contest specialization* (see Figure 1.1). In competitive settings with high contest specialization, ability or expertise is the key determinant of solver performance, such as a scientific contest offered on InnoCentive.com or, at an extreme, competing with the best nuclear physicists in the world to develop the first atomic bomb, as it was the case in the Manhattan Project (Lenfle and Loch, 2010). Conversely, solver output depends mainly on the amount of effort exerted in settings with low contest specialization, such as choosing the best Chinese interpreter for a business meeting on UpWork.com.

Second, solver participation in a contest cannot be guaranteed because each solver faces a non-negligible participation cost. For instance, solver participation is voluntary for innovation contests on InnoCentive.com or at logo design contests found on 99designs.com. At an extreme, one can also conceptualize a ride-sharing platform as a contest in which independent contractors can self-select whether or not to work and compete for available demand. This is a primarily effort-driven contest with low contest specialization, as everyone who is capable of driving and owns a car can become an Uber driver.

In studying innovation contests, the existing literature largely finds that a winner-takes-all (WTA) award scheme is optimal. This is due to the fact that they either consider contests in which effort alone determines performance, or they neglect the strategic participation decision

Low Contest specialization **High**
Primarily effort-driven outcome Primarily ability-driven outcome

Figure 1.1 Different structural properties of contests.

	Effort-only contests	Ability and Effort contests
Participation is endogenous	Kalra and Shi (2001)	This paper
Everyone participates	Moldovanu and Sela (2001) Moldovanu, Sela and Shi (2007)	Terwiesch and Xu (2008) Körpeoğlu and Cho (2017)

Figure 1.2 Brief contest literature taxonomy.

of solvers and analyze only the case when solvers have zero participation cost. Moldovanu and Sela (2001) consider a setting where effort (alone) is observable by the seeker and they show that WTA is optimal when the cost of effort is linear or concave. In contrast, in this chapter, solver effort is unobservable and the output of each solver is a function of both effort and (privately known) ability. In addition, the solvers in our setting face a strictly positive participation cost and a linear cost of effort. Our setting leads to multiple awards rather than WTA. Kalra and Shi (2001) do consider non-negative participation costs and account for unobservable effort, but in their model the solvers are homogeneous in terms of ability. In a recent paper Nittala and Krishnan (2016) study an internal innovation contest and allow the seeker to face a non-negligible participation cost; we complement this work focusing on solvers' cost to participate. Because all solvers are homogeneous, either all of them participate or none of them participate. The result is, once again, a WTA award scheme. Our setting allows for heterogeneity among solvers due to their intrinsic ability. Depending on the award structure offered, only a subset of the solver population chooses to participate and exert effort. Terwiesch and Xu (2008) and Körpeoğlu and Cho (2017) also study contests with ability and effort, however, they do so normalizing solver participation cost to zero, leading to a WTA award scheme. The seemingly innocuous assumption regarding participation cost fails to address solvers' endogenous participation choice, which is central in large crowdsourcing platforms. As a consequence, all solvers enter the contest and the seeker does not face a trade-off between how many solvers and of what skill level to encourage to enter, and how to motivate those solvers that do enter to work hard. Conversely, we demonstrate how the trade-off between participation and effort forces the seeker to provide multiple awards.

The purpose of this chapter is to find the incentive structure design needed to balance the strategic participation and effort choices of the solvers. We provide a theoretical basis for the prevalence of offering multiple awards in practical settings where solver participation is voluntary. To do so, we develop a game-theoretic model of incomplete information in which each solver first chooses whether to participate and then how much effort to exert in the innovation contest. A solver's intrinsic ability (e.g. skill or expertise level) and effort together determine his output. If a solver exerts more effort, he increases the chances that he will receive an award, but effort is costly and it is more costly for solvers with lower ability. Further, in order to make

a rational participation decision, each participating solver has to cover a fixed participation cost to enter the contest. This is a critical component of our model, and one of the key differences between our work and the existing research in this area.

Our results offer guidelines for innovation contest designers depending on their objectives and whether the characteristics of the innovation contest require greater incentives for participation or effort. Specifically, our analysis yields three main results. First, we prove that any monotone reward allocation induces a (unique) threshold ability participation strategy for the solvers in which their equilibrium performance is strictly increasing in ability. Interestingly, while solver equilibrium performance is strictly increasing in ability, solver equilibrium effort exhibits a non-monotone behavior (i.e. high-skilled solvers substitute ability for effort, whereas effort complements ability for low-skilled solvers). As a consequence, in contests which require a sufficiently high degree of specialization, the solvers substitute ability for effort to the extent that all participating solvers exert zero effort in equilibrium.

Second, we find that the contest designer can mitigate these adverse effects by providing multiple awards. In particular, we show that the celebrated winner-takes-all result of Moldovanu and Sela (2001), which maximizes total effort, extends to effort-only contests in which solver participation is voluntary. In contrast, when the performance of a submitted solution is affected by a combination of solver ability and effort, the seeker may find it optimal to offer multiple awards. All of our equilibrium results are distribution-free. In our model we impose a linear cost of effort and analyze it using structural properties of order statistics theory. The presence of a convex cost of effort would only strengthen our arguments in favor of offering multiple awards. As it turns out, offering a single award can be potentially quite damaging; we present a numerical example in which a winner-takes-all budget allocation performs 20% *worse* than the optimal multiple award allocation.

Third, our analysis shows that offering multiple awards is beneficial in a general objective of optimizing a weighted combination of the total performance of the best *candidate solutions*, i.e. when the contest designer is interested in the performance of more than one solution (but not necessarily all of them). The optimal allocation of multiple awards balances a novel trade-off to maintain a desirable participation level while at the same time provides incentives for participating solvers to exert effort. Our previously obtained results remain robust in this general objective. The optimal award allocation contains no more than an upper bound of awards, which depends on the structural properties of the contest and the candidate solutions that the seeker optimizes over.

1.2 An innovation contest model

An innovation contest *seeker* ("she") has a fixed budget R and elicits innovation from outside solvers with population size N. Solver participation is voluntary and solvers face a fixed set-up cost to participate $c_p > 0$. To establish that solvers' voluntary participation decision makes seeker offer multiple awards, we assume that the seeker splits her budget evenly into m equal rewards according to a subset of allocations termed *Multiple-Winners* (MW), of which *Winner-Takes-All* (WTA) is a special case with $m := 1$. The jth reward has the value $R_j := \frac{R}{m}$.

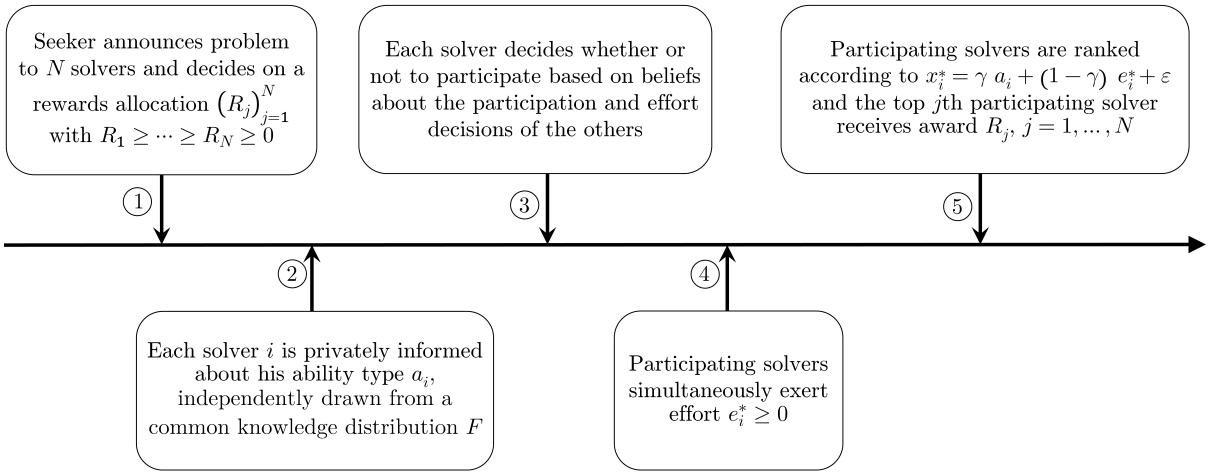


Figure 1.3 Sequence of events in an innovation contest game.

A MW format captures the key trade-off most seekers face between offering many rewards of smaller value versus being more selective and distribute fewer rewards of higher value[2]. To accommodate solver endogenous participation, we allow the reward structure to be contingent on the actual number of participants[3]. That is, given a chosen allocation $(R_j)_{j=1}^N$ and depending on the realized number of participating solvers $n \in \{0, 1, \ldots, N\}$, the seeker may not expend her entire budget, i.e. $\sum_{j=1}^N R_j \leq R$.

One of the benefits of crowdsourcing innovation is that outside solvers exhibit a significant heterogeneity in terms of their expertise. Each solver is privately informed about his own *ability* (type) a_i. Abilities are drawn independently of each other from a common knowledge distribution F which is strictly increasing on its support $[a_0, 1]$, where $0 < a_0 < 1$. Knowing his ability a_i, solver i exerts *effort* $e_i \geq 0$ and incurs a linear cost $c(a_i, e_i) := \frac{e_i}{a_i}$. The latter formulation of solvers' cost follows Moldovanu and Sela (2001) and implies that solvers exhibit quasi-linear preferences, as it is standard in the mechanism design literature. From the perspective of solver i, a_i is a constant and our simple linear cost function suggests that solvers of higher ability have lower marginal cost of effort. As noted by Moldovanu and Sela (2001) the key assumption here is the separability of solvers' ability and effort.

The seeker observes the performance of the solutions of the participating solvers (but not their ability and efforts separately) and ranks them in a relative order according to her subjective taste which is not known a priori by the solvers. We model the *performance* (or output level) of solver i as

$$x_i := \gamma\, a_i + (1-\gamma)\; e_i + \varepsilon,$$

where ε is the random shock realization of seeker's subjective taste of a submitted solution[4],

[2] Although seemingly limiting, the preceding assumption on the reward structure is made *only* for the easy of exposition and to focus on the latter key trade-off; we demonstrate the robustness of our results by solving the general combinatorial problem of the seeker allocating a weakly decreasing reward allocation in §D.4.

[3] We note that when nobody enters, the seeker keeps the budget. Similarly, when one solver enters, he exerts zero effort and wins the reward $R_1$with certainty.

[4] In line with the innovation and R&D literature (Terwiesch and Xu, 2008; Erat and Krishnan, 2012; Mihm and Schlapp, 2015; Körpeoğlu and Cho, 2017), this implies that the seeker can not fully specify ex ante her precise evaluation criteria. Since the seeker evaluates submitted solutions in a single round, it is natural to assume

and $\gamma \in [0,1]$ is a structural (exogenous) characteristic of each contest we refer to as *contest specialization*. The contest specialization captures the degree to which the solvers can substitute effort for ability to determine the output of their solution. Indeed, in a variety of contests, the performance of a participating solver is affected to a certain degree by a "fixed effect" due to his intrinsic ability irrespective of his effort choice. Note that the special cases of $\gamma = 0$ and $\gamma = 0.5$ result in the observable effort model of Moldovanu and Sela (2001) and the performance function of Terwiesch and Xu (2008) with a logarithmic effort transformation, respectively.

In order to make rational participation and effort decisions, a solver faces two kinds of uncertainties: he is unsure of the *number* of solvers who will choose to participate, as well as of his own *rank-order* among any participants. Following Stouras et al. (2016) we model this "strategic uncertainty" assuming that a solver forms (ex ante) *beliefs* about the anticipated participation and effort actions of the other solvers, conditioned on his own action. In a *symmetric Bayes-Nash equilibrium* (BNE) we require solver beliefs to be symmetric and self-confirming. In particular, a solver first determines his participation probability by making a conjecture on the rest (uncertain) number of participants to participate with (ex ante) participation probability $\tilde{p}$. In a BNE we require $\tilde{p}$ to equal the actual (ex post) participation probability p^*. Upon entry, solvers simultaneously determine their efforts without observing the number or the types of the participants. We describe the timing of our static game in Figure 1.3.

The utility of a solver is affected by his individual participation and effort decisions, as well as by the decisions of the other solvers. The participating solver with the highest performance among the *participants* wins the reward R_1. The participating solver with the second highest performance among the participants wins the reward R_2, and so on until all rewards are allocated. When nobody enters, the seeker keeps the budget. Similarly, when one solver enters, he exerts zero effort and wins the reward R_1 with certainty. That is, the *utility* of a participating solver i is either $R_j - \frac{e_i}{a_i} - c_p$ if he wins reward R_j, or $-\frac{e_i}{a_i} - c_p$ if he does not win a reward. Hence, based on belief $\tilde{p}$ about solvers' participation probability, solver i participates and competes with any other participating solvers, if and only if, his *expected utility* $u\left(a_i,\, e_i^*;\, \tilde{p}\right)$ from doing so covers the participation cost $c_p > 0$ (IR constraint), where $e_i^* = \arg\max_{e_i \geq 0} u\left(a_i,\, e_i;\, \tilde{p}\right)$ (IC constraint) and $\tilde{p} = p^*$ (self-confirming equilibrium (SCE) beliefs condition).

The risk neutral seeker values multiple solutions $k \in \{1, \ldots, N\}$ from participating solvers and her objective is defined as[5]

$$\max_{1 \leq m \leq N} \Pi_k\left(m;\, p^*\right) = \mathbb{E}\left[\sum_{i=1}^{k} w_i \left\{x_i^*\left(\mathcal{A}_i;\, m,\, p^*\right) \text{ is ranked } i\text{th out of } N\right\} \cdot \mathbb{1}_{\{i \text{ participates}\}}\left(m;\, p^*\right)\right] \tag{1.1}$$

for exogenously specified weights $w_1 \geq w_2 \geq \ldots \geq w_k > 0$, where the expectation operator is taken over any sources of randomness. We note that the chosen reward mechanism moderated by seeker's choice of m affects the functional form of effort exerted in equilibrium, as well as,

that seeker's taste is a *common* random shock across solvers (symmetric noise).

[5]Since the ability type of solver i is private information to him, his equilibrium output is a random variable $x_i^*\left(\mathcal{A}_i;\, m,\, \beta^*\right)$ for the seeker. We denote any random variables with calligraphic symbols to distinguish them from any exogenous variables denoted with capital letters. Since the ability distribution F is common knowledge and based on self-confirming beliefs, The seeker can correctly calculate the distribution of the participating agents and the expected value of any order-statistics derived from the solvers' ability $\mathcal{A}$.

the participation decision of the solvers. This further influences seeker's objective since only the participating solvers generate output due to the indicator function in (1.1). We provide a summary of our notation in §D.1.

1.3 Solvers' equilibrium

In this section, we analyze the strategic behavior of the solvers in equilibrium focusing on symmetric pure strategies. We first show that a seeker who offers the same award to any solver who chooses to participate would either attract all solvers, or none in a symmetric equilibrium.

Lemma 1 (Restrictions on seeker's allocation)**.** *(a) Suppose that the seeker decides to allocate N equal rewards (i.e. $m = N$). Then, there exist a unique symmetric equilibrium in which solvers participate with probability $p^* = \min\left\{\max\left\{0, \frac{R-c_p}{(N-1)c_p}\right\}, 1\right\}$. In pure and symmetric strategies either all, or none of the N solvers participate and exert zero effort.*

(b) Assume that the seeker allocates m rewards (of any size) and set

$$\overline{m} := \max\left\{n \in \{1,\dots,N\} : \frac{R}{n} \geq c_p\right\} = \max\left\{1, \left\lceil \frac{R}{c_p} \right\rceil - 1\right\} \qquad \text{(Upper bound on awards)}$$

In a symmetric pure equilibrium: (i) If $m > \overline{m}$, then no solvers participate. (ii) If $m = \overline{m}$, then all solvers participate and exert zero effort. (iii) If $m < \overline{m}$, then $\mathcal{N} \sim$ Binomial(N, p) solvers participate with the same probability $p < 1$.

Lemma 1 restricts any weakly monotone allocation of rewards of the seeker. In particular, splitting the total budget among the entire solver population results in a unique symmetric equilibrium in which solvers participate with a probability p. In a symmetric equilibrium the ex ante number of participating solvers is a non-negative random variable that follows the Binomial (N, p) distribution. That is, we distinguish between the stochastic number of *participating* solvers denoted with $\mathcal{N}$, and the number of potential solvers (or solver population size) N which is a deterministic quantity and common knowledge. In order to avoid corner cases and focus on interior solutions, we require that the seeker cannot set N awards of positive value, so that the solvers participate with a probability that is strictly less than one.

In addition, Lemma 1(b) provides an upper bound on the number of rewards to induce a subset of solver population to participate and exert non-zero effort. Importantly, such an upper bound on the number of rewards to allocate holds for *any* weakly monotone allocation of the budget to participating solvers. We note that this is typically satisfied in large innovation contests in practice. For instance, consider an innovation tournament at InnoCentive in which seeker's satisfies $R \leq 5\,c_p$. Then, as long as the size of solvers' population satisfies $N \geq 6$, Lemma 1 suggests that the seeker can not allocate more than $\overline{m} = 4$ rewards, as by doing so will result in no participation from the solvers. Further, to induce some solvers to participate and expend non-zero effort, Lemma 1(b) shows that the seeker should allocate at most three rewards in this case.

Next, we characterize solvers' participation and effort strategy conditioned on MW budget allocations with a number of $m \in \{1,\dots,\overline{m}\}$ rewards pre-announced by the seeker. The analysis

of the general, weakly monotone allocation is qualitatively similar and is delegated to §D.4. We denote by $F_{N-m:\,N-1}(\cdot)$ and $f_{N-m:\,N-1}(\cdot)$ the $(N-m)$th lowest out of $N-1$ order statistics CDF and PDF of the ability distribution $F(\cdot)$ respectively. We present our results in two separate theorems by first describing solvers' endogenous participation strategy and then solvers' strategic effort decision.

Theorem 1 (Solvers' participation strategy)**.** *Suppose that the seeker has chosen to allocate her budget R into m rewards, where $m \in \{1, \ldots, \overline{m}\}$. Then, there exists a unique $a_{min} \in [a_0, 1]$ that solves*

$$\frac{R}{m} \cdot F_{N-m:\,N-1}(a_{min}) = c_p \tag{1.2}$$

such that only $\mathcal{N}$ solvers with ability $a_i \geq a_{min}$ participate with ex ante probability $p^*(N) := 1 - F(a_{min}(N))$.

As we show in the proof of Theorem 1, for any allocation of the available budget chosen by the seeker the expected utility of a solver is strictly increasing in his ability. That is, if a solver of a given ability finds it rational to participate, all solvers of higher ability participate as well. As a consequence, solvers' participation strategy is characterized by a threshold $a_{min} \in [a_0, 1]$ that defines the "marginal solver" who is indifferent between participating and not paying the cost to participate. Naturally, the existence of a unique and symmetric BNE implies that the solvers participate by choosing a probability $p^* = 1 - F(a_{min})$. Due to symmetry, the participation probability of a solver does not depend on his ability. Further, we note that solvers' participation probability p^* depends on seeker's reward allocation and budget, as well as solver population size, participation cost and ability distribution. As we show below, these dependencies are critical in order to understand the effect of solver participation on the nature of competition among participants.

Next, we analyze solvers' effort decision conditional on participation. We establish that the solvers have a unique, symmetric and pure effort strategy that crucially depends on contest specialization.

Theorem 2 (Solvers' equilibrium effort)**.** *Let $a_{min}(m)$ be the induced ex-ante participation threshold of the solvers and define $\hat{\gamma} := \left(1 - \frac{1}{a_{min}(m) \cdot R}\right)^+$. Then, for any fixed contest specialization $\gamma \in (\hat{\gamma}, 1]$: $e^*(a; \gamma) = 0$ for all participating solvers with ability $a \in [a_{min}(m), 1]$.*

When $\gamma \in [0, \hat{\gamma}]$ a solver with ability a_i exerts equilibrium effort

$$e^*(a_i; \gamma, m) = \begin{cases} \frac{\gamma}{1-\gamma}\,(a_{min}(m) - a_i) + \frac{R}{m} \cdot \int_{a_{min}(m)}^{a_i} x \cdot f_{N-m:\,N-1}(x)\,dx > 0, & a_i \geq a_{min}(m) \\ 0, & a_i < a_{min}(m) \end{cases} \tag{1.3}$$

and (relative) equilibrium performance

$$x^*(a_i) = \gamma\, a_{min}(m) + (1-\gamma)\,\frac{R}{m} \cdot \int_{a_{min}(m)}^{a_i} x \cdot f_{N-m:\,N-1}(x)\,dx \tag{1.4}$$

At the core of our conceptual contribution is that participating solvers exert effort to compete only with the rest of the *participants* (as opposed to solver total population). Prior literature that builds on the model of Moldovanu and Sela (2001) analyzed contest settings either by

specifying an exogenously fixed number of participants or by normalizing solvers' participation cost to zero (Körpeoğlu and Cho, 2017). The latter neglects the moderating effect of the solver voluntary participation choice on solver equilibrium effort decision. In contrast, we explicitly account for the endogenous participation choice of the solvers. The closed-form expressions of solver equilibrium effort (eq. 1.3) and performance (eq. 1.4) are rather intuitive and connect the offered budget allocation with the probability to achieve one of the multiple possible rankings through the order statistics distribution of ability. In the special case of the WTA contest design (i.e. $m = 1$), our equilibrium effort expression (1.3) simplifies to the equilibrium effort and performance of Proposition 2 of Körpeoğlu and Cho (2017) by a log-transformation of their ability distribution.

Theorem 2 proves that if the nature of an innovation contest is sufficiently specialized such that solver performance is mainly driven by solvers' ability, solvers exert minimal effort. Depending on seeker's budget, there exists a contest characterized by a "critical" contest specialization $\hat{\gamma}$ defined in Theorem 2 which suggests an explicit lower bound on seeker's budget in order to guarantee non-zero effort by participating solvers. In simple terms, the higher the ability required for a given task manifested by a higher contest specialization, the larger is the amount of the budget required to sustain the competition among the solvers. Coupled with the Lemma 1, we assume that seeker's budget satisfies the following lower and upper bounds.

Proposition 2 (Budget condition)**.** *Solvers participate with probability $p^* \in (0, 1)$, if and only if, seeker budget satisfies*

$$\max\left\{\frac{1}{a_0},\, c_p\right\} < R < N\, c_p, \qquad \text{(Budget Condition)}$$

which also guarantees that participating solvers exert strictly positive equilibrium effort.

Proposition 2 derives a necessary and sufficient condition on seeker's budget for agents to participate with a non-trivial probability which also ensures that solvers exert strictly positive effort in equilibrium. The Budget Condition summarizes the following three effects. First, the budget R should exceed solver participation cost c_p, as otherwise no solver finds it rational to participate in equilibrium.

Second, we require seeker's budget to be limited ($R < N\, c_p$), in order to guarantee that only a *subset* of the solvers' entire population can potentially participate. We note that similar upper bound on budget is imposed in the Corollary 2 of Erat and Krishnan (2012) in a related setting. Due to Lemma 1(a) we have that solver population participates with zero probability. How many solvers (and of which ability) the seeker chooses to motivate to participate is then a non-trivial question. When the Budget Condition is violated, the entire population of N solvers participate. The latter is not aligned with what is observed in practice in large-scale crowdsourcing platforms such as Tongal.com (see the recent empirical evidence of Table 2 of Kireyev, 2016).

Third, a sufficient (but by no means necessary) condition to guarantee that participating solvers exert strictly positive equilibrium effort is that the marginal cost of effort is sufficiently low, or that $\frac{1}{a_i} < R$, for each agent i who participates. Intuitively, the Budget Condition ensures that no solver relies entirely to his ability when participating in the contest. That is,

the "critical" contest specialization $\hat{\gamma}$ is one. As we demonstrate in Theorem 15 in §D.4, similar bounds can be obtained for innovation contests with a general reward structure (rather than the special case of MW contests).

This structural outcome of our model is consistent with the anecdotal evidence that high skilled solvers exert minimal effort while they still manage to maintain high outputs. Additionally, observe that for a given threshold participation ability a_{min}, the first term in solvers' effort (1.3) is negative with respect to their ability, while the second term is strictly increasing in ability. The latter implies that solvers' equilibrium effort is in general *non-monotone* in their skill level. Consider the case of some MBA students seeking a consulting interview. Applicants with more expertise can achieve higher performance and still devote less time to prepare.

In the special case of the WTA contest design (i.e. $m = 1$), expression (1.3) simplifies to the expression (17) of Körpeoğlu and Cho (2017) by setting their reward $A_2 := 0$ and applying a log-transformation in the ability distribution. While Körpeoğlu and Cho (2017) contain a graphical illustration that is aligned with the spirit of our result, the underlying mechanism of our settings differ: our performance function is linear in solvers' effort. Accounting for a sufficiently concave function of the effort (such as the natural logarithm) when ranking the solvers, in combination with an ability distribution that is strictly log-concave (such as the Gumbel distribution considered by Terwiesch and Xu (2008) and Körpeoğlu and Cho (2017)) may drive a non-monotonic effort choice of the solvers with respect to their skill level. Interestingly, we show that this behavior is sustained even when output is a *linear* function of effort (see Figure 1.4(C) for an illustration).

Understanding the behavior of the marginal solver and its dependencies with the exogenous parameters of the contest are critical in identifying potential underlying moderating mechanisms of solvers' strategic choices. The next result provides useful comparative statics insights on seeker's choices, on the exogenous characteristics of the contest that are outside the control of the seeker to a large extent, as well as on the characteristics of the competing solvers.

Corollary 1. *(a)* Sensitivity to seeker choices: *The ability threshold a_{min} is non-monotone in seeker's choice of the number of awards m with a unique minimum. In particular, there exists a unique m_0^* such that $a_{min}(m) > a_{min}(m_0^*)$ for all $m < m_0^*$, and $a_{min}(m_0^*) > a_{min}(m)$ for all $m > m_0^*$.*

(b) Sensitivity to solver characteristics: *The ability threshold a_{min} is weakly increasing in solvers' population N and participation cost c_p. Further, a_{min} is globally the same for all solvers.*

(c) Sensitivity to contest characteristics: *The ability threshold a_{min} does not depend on the contest specialization γ and is a function of the order statistics distribution of ability in the population.*

Extending the number of awards increases the probability of winning a given award, whereas the value of each award diminishes. One may think that an increase in the number of awards would always induce more solvers to participate. However, in Corollary 1(a) we prove that more awards only attract more participants up to a "tipping point" beyond which a higher number of awards causes the participation levels to drop. In addition, Corollary 1(b) shows

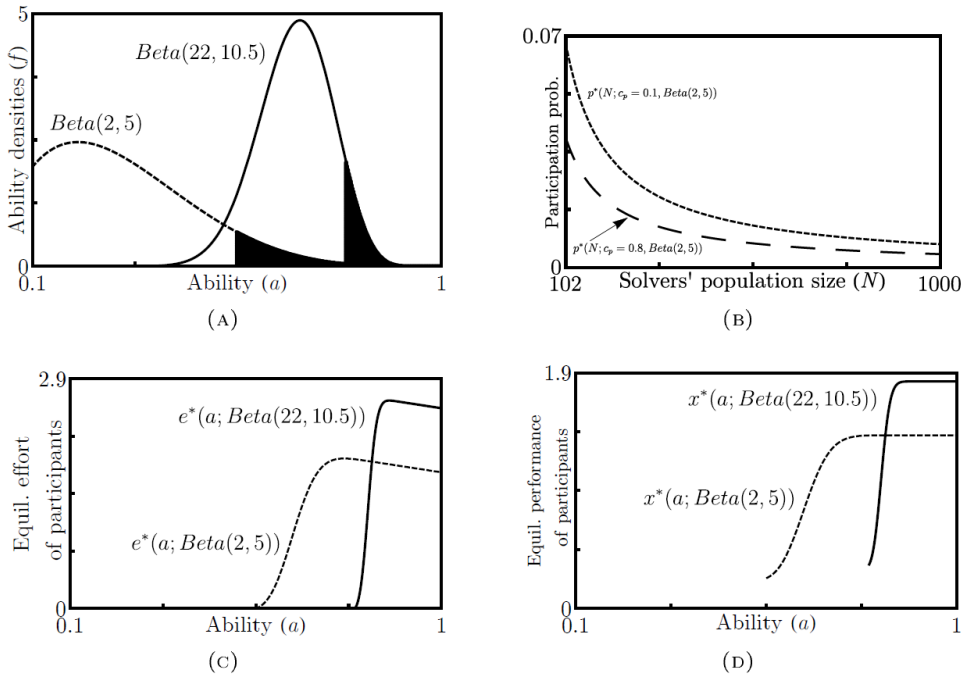


Figure 1.4 Suppose that $R = \$10$, $c_p = \$0.1$, $m = 3$ awards, $N = 102$ agents, and $a_0 = 0.1$. **(A)** The mass of participating agents (cf. shaded area) for the Beta (2, 5) and Beta (22, 10.5) ability distributions. **(B)** Probability to participate in equilibrium (p^*) as a function of agent population size (N). We use the Beta (2, 5) ability distribution and we plot for $c_p = \$0.1$ and $c_p = \$0.8$. We plot the unimodal equilibrium effort functions on Panel **(C)** and the strictly increasing equilibrium performance functions on Panel **(D)** for contest specialization $\gamma = 44\%$ and Beta (2, 5) and Beta (22, 10.5) ability distributions, respectively.

that such a threshold is the same for all solvers and does not depend on their ability realization, or properties of the contest.

However, for a given award structure chosen by the seeker, the specific characteristics of the competition do not directly affect solvers' participation choice (Corollary 1(c)), but are indirectly present through the subtle dependence of the contest specialization on the lower bound of awards to set, as discussed in Lemma 1. Competing with a large population of potential solvers would decrease the participation probability of each solver. Indeed, the ability threshold of the "marginal solver" can not decrease as the population increases. The reverse intuition holds for the relationship between the ability of the marginal solver and the participation cost; the larger the cost to participate, the higher the ability participation threshold, i.e. the lower the expected number of submitted solutions. Further, the exogenous attributes of a given contest captured by the value of the contest specialization negatively affect solvers' choice of effort, as Theorem 2 suggests.

The dependence of the marginal solver on the characteristics of the ability distribution implied by Corollary 1(c) is more involved. Our results are *distribution-free* and hold for any ability distribution which is strictly increasing on its support. To study the effect of the shape of the ability distribution, we consider the Beta distribution. It is known that the density Beta (α, β) is symmetric when $\alpha = \beta$; it is unimodal when $\alpha > 1$ and $\beta > 1$ with a positive skew (right-

tailed) when $\alpha < \beta$; and it is unimodal with a negative skew (left-tailed) when $\alpha > \beta$. Further, when $\alpha = \beta = 0.5$ the Beta distribution is identical to the standard Uniform distribution.

We summarize three key managerial insights obtained so far in Figure 1.4. First, the mass of participating solvers depends on the size of the population, the ability distribution and the reward allocation. Suppose that the seeker awards three equal rewards (i.e. $m = 3$) and that high-skilled solvers are rare. This models a contest that attracted a very diverse pool of solvers from a right-tailed ability distribution. Using simulation we illustrate the mass of participating solvers by the shaded area of the left-tailed Beta (22, 10.5) density in Figure 1.4(A) and by the shaded area of the right-tailed Beta (2, 5) density respectively. The marginal participating solver from the Beta (22, 10.5) density has higher ability, thus the screening effect is more pronounced for left-tailed distributions.

Second, as the population of potential contestants grows larger, the chance that a solver wins a reward evaporates. This significantly decreases a solver's participation probability. Then, it is intuitive that a higher participation cost decreases his probability to participate (see Figure 1.4(B) for an illustration).

Third, participating solvers' equilibrium effort is non-monotone with a single peak. In Figure 1.4 (C) we plot the equilibrium effort as a function of ability for the right-tailed Beta (2, 5) and left-tailed Beta (22, 10.5) density respectively. We observe that there is a unique solver we refer to as the "hardest worker"; the worker who defines the single peak in Figure 1.4(C). Hence, two regions arise in equilibrium: those solvers with ability lower than the hardest worker and those above him. In the first region, the ability and effort behave as *complements* for solvers with lower ability than the "hardest worker", and as *substitutes* otherwise. Qualitatively, the lower the skewness of the ability distribution, the larger is the region where ability and effort are substitutes. That is, the peak of the effort is attained at a higher ability as the skewness of the ability distribution increases. Lastly, we note that the equilibrium performance is strictly increasing in ability as shown in the proof of Theorem 2. This is also illustrated in the plot of the equilibrium performance as a function of ability in Figure 1.4(D) for the right-tailed Beta (2, 5) and left-tailed Beta (22, 10.5) density respectively.

1.4 Seeker's problem

The results from the previous section emphasized the role of reward allocation on solvers' strategic participation and effort choices. In this section, we analyze how the seeker should optimally manage an innovation contest with endogenous participation, given a specific degree of contest specialization.

To build intuition, we first consider a special case of the seeker's problem. Specifically, we investigate how the optimal MW rewards should be determined by the seeker to maximize the total output of the *participating* solvers. Subsequently, we examine the optimal reward structure to maximize a weighted combination of the top performers that participate, as well as its dependency on the parameters of the contest.

1.4-1 Maximizing the expected total performance of the participants

In various practical settings the contest seeker derives benefit from the collective output created by all solvers who self-selected to participate, rather than from the solvers of any special subgroup of the population. In such settings the seeker wishes to determine the award allocation that induces all participating solvers of *any* performance rank-order to generate the maximum output. Normalizing all weights $w_i = 1$ for all positions $i = 1, \ldots, N$, we derive a closed form expression that describes the seeker's objective (1.1) as a function of the number of awards m chosen by the seeker.

Lemma 2 (Total performance of participants)**.** *Assume that $w_i = 1$ for all $i = 1, \ldots, N$. The total expected performance of all participating solvers in equilibrium is given by*

$$\Pi_N(m;\, \gamma) = \begin{cases} N\,\gamma\, a_{min}(m) \cdot (1 - F(a_{min}(m))) + (1-\gamma)\; R \cdot \mathbb{E}[m,\, N;\, a_{min}(m)], & \gamma \in [0,\, 1) \\ N \cdot \mathbb{E}[\mathcal{A} \,|\, \mathcal{A} \geq a_{min}(m)], & \gamma = 1 \end{cases} \tag{1.5}$$

where $\mathbb{E}[m,\, N;\, a_{min}(m)] := \int_{a_{min}(m)}^{1} x \cdot f_{m:\,N-1}(x)\; dx$.

The seeker's decision on the number of rewards to offer impacts her objective through the following three interconnected ways. The first, which we refer to as the "effort effect", reflects the effect of seeker's choice of rewards on the expected total effort elicited by the solvers. We note that the "effort effect" has been a central focus of the contest literature after Moldovanu and Sela (2001). The second, which we refer to as the "screening effect", captures the effect of seeker's allocation on the *support* of the total expected effort exerted by participating solvers. The third, which we refer to as the "participation effect", shows the effect of seeker's choice of rewards on the marginal solver who essentially defines the number and the skills of solvers who choose to participate. The latter two novel effects we identify are due to an incentive misalignment between the solvers' individual preferences and the objectives of the seeker. Overall, the optimal allocation balances all these three effects.

Our results provide a proof using order statistics that supports the suggestion of Moldovanu and Sela that their WTA result continues to hold when solvers face a strictly positive participation cost $c_p > 0$ (see Moldovanu and Sela (2001), pp.550-551). Observe that if the solvers are ranked purely based on their choice of effort, i.e. $\gamma = 0$ then seeker's objective takes the form: $\Pi_N(m) = R \cdot \mathbb{E}[N - m,\, N;\, a_{min}]$. If we set $m := 1$ and substitute $V_2 = 0$ and $V_1 = R$, then our closed form expression (1.5) agrees with formula (4) on p.547 of Moldovanu and Sela (2001) by re-interpreting the quantities involved.

Theorem 3. *Assume that solvers' participation cost satisfies $c_p > 0$.*

(a) If the contest specialization is zero ($\gamma = 0$), then the WTA allocation is optimal.

(b) For a non-zero contest specialization $\gamma \in (0,\, 1]$ the optimal allocation contains m^* awards, where $m^* \in \{1, \ldots,\, m_0^*\}$ and m_0^* is given by Corollary 1(a). In particular, the WTA allocation is not always optimal.

Theorem 3 shows that the most informative, finely hierarchical reward allocation is never desirable for the seeker. In particular, allocating more than m_0^* awards *hurts* the seeker's objective. To see this, recall from Corollary 1(a) that m_0^* is the optimal number of awards to minimize the induced threshold a_{min}; hence, allocating more than m_0^* is harmful to the seeker's

objective as it reduces *participation*. Additionally, as we show in Theorem 3(a) granting one award to the top (WTA scheme) maximizes the total expected *effort* of any participating solvers. Overall, offering more than m_0^* awards is detrimental for both the participation and the effort incentives of the solvers, and hence it is never desirable by the seeker.

The beneficial effect of offering multiple awards in order to maximize the total output has been identified by the previous literature but the rationale behind it is fundamentally different. That is, once we recognize that not all solvers participate, Theorem 3(b) shows that multiple awards can be profitable even in the conservative case of *linear* cost of effort, and irrespective of the convexity of the distribution of ability. In contrast, Moldovanu and Sela (2001) rely on the special case where everyone participates and show that when the cost of effort is "sufficiently convex", multiple awards are optimal. In a follow-up paper and based on another convexity argument on the ability distribution, Moldovanu et al. (2007) show that the most hierarchical reward structure can be optimal, when all solvers participate. We note that the presence of a non-linear costs of effort in our setting would only strengthen our arguments in favor of multiple awards.

Recently, Megidish and Sela (2013) argue that when a minimal effort is required in order to compete, the expected total effort under a WTA is dominated by a "random contest" in which the entire budget is equally allocated among all the participants. We generalize this intuition and characterize the optimal budget allocation when solvers are ranked based on a convex combination of ability and effort. Next, we provide a simple example with a strictly positive participation cost in which WTA is not optimal, due to the "participation" and "screening" effects of the strategic solvers.

Example. Consider an innovation contest with a population of $N = 102$ potential solvers whose ability follows the Unif $([0.1, 1])$ distribution. Due to the fact that an accurate estimate of the participation cost of the solvers is challenging, assume that the ratio of the seeker's budget to solvers' participation cost is $\frac{R}{c_p} = 66.66$. Suppose also that the nature of the contest implies a contest specialization of $\gamma = 30\%$. As Figure 1.5(A) illustrates, choosing a WTA budget allocation and ignoring the "participation" and "screening" effects results in a substantial decrease of *at least* 19% in seeker's objective (note that this is a conservative estimate since we are optimizing over MW allocations; see §D.4 for the solution to the general combinatorial problem).

In lieu of the above, we investigate how the seeker's optimal choice of awards depends on the exogenous parameters of the innovation contest.

Theorem 4 (Comparative statics)**.** *(a) A seeker having a larger budget R should (weakly) increase the optimal number of awards. That is, all else being equal, $m^*(R)$ is weakly increasing.*

(b) A seeker of an innovation contest with higher contest specialization γ should (weakly) increase the optimal number of awards. That is, all else being equal, $m^*(\gamma)$ is weakly increasing.

(c) As solvers' participation cost c_p increases, the seeker should (weakly) decrease the optimal number of awards. That is, all else being equal, $m^*(c_p)$ is weakly decreasing.

(d) As solvers' population N increases, the seeker should (weakly) increase the optimal number of awards. That is, all else being equal, $m^*(N)$ is weakly increasing.

Claim (a) states that a seeker with a larger budget would optimally announce multiple awards,

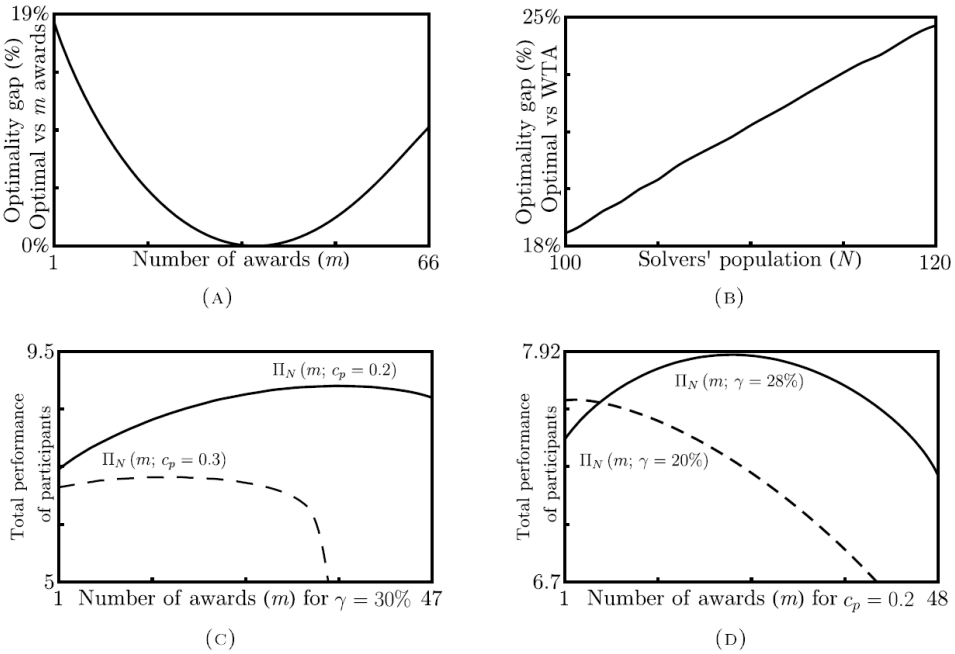


Figure 1.5 Suppose that $R = \$10$, $N = 102$ agents and $a_0 = 0.1$. For $c_p = \$0.15$ and $\gamma = 30\%$ we plot the optimality gap ratio: $\frac{\Pi_N(m^*)-\Pi_N(m)}{\Pi_N(m^*)} \cdot 100\%$ as a function of awards (m), and solver population size (N) in Panel **(A)** and **(B)** respectively. We plot the total performance of participating solvers as a function of awards for contest specialization $\gamma = 30\%$ and various participation costs in Panel **(C)**, and for solver participation cost $c_p = 0.2$ and various contest specializations in Panel **(D)**.

which shrinks the value of each individual award as the budget is fixed. Intuitively, this increases the probability of a solver to receive an award, which effectively decreases their participation cost in the contest. This naturally leads to an increase in participation to the benefit of the seeker. The opposite effect holds true if the solvers face a higher participation cost. In particular, an increase in the solvers' participation cost can be outweighed with a suitable increase in the resources invested in the contest. Figure 1.5(C) graphically illustrates this result.

Similarly, Claim (b) shows that, all else being equal, the seeker should weakly increase the optimal number of awards subject to an increase in contest specialization. In lieu of the optimality of the WTA scheme shown in Theorem 3(a) when effort is all that matters in solvers' rankings (i.e. when γ is in a neighborhood of zero) and due to the continuity of the seeker's objective in γ we would expect that *few* awards, if not WTA, to be optimal. In contrast, for large values of the contest specialization factor we would expect a large number of awards to be optimal due to Theorem 4(b). Hence, by continuity it is intuitive that as γ increases, the weight on the "participation effect" increases while at the same time the weight on the "effort effect" decreases, which force the optimal number of awards to (weakly) increase. We summarize the insights in Figure 1.5(D).

Previous literature has demonstrated that an increase in the number of potentially competing solvers is beneficial to the total expected effort at any optimal reward allocation (Moldovanu et al., 2007). Due to the "participation-related terms" in the seeker's objective (1.5) and the

zero-sum nature of the innovation contest game, it is not clear a priori whether an increase in the value of an award or the allocation of the budget across more awards would be a better strategy for the seeker. Claim (d) of Theorem 4 proves that a seeker faced with a larger pool of potential solvers should optimally prefer to reduce the competition among solvers. To do this, a seeker should offer the same or less rewards. In addition, this implies that the gap between the optimal and the WTA allocation increases in the population of the solvers. We illustrate this result in Figure 1.5(B).

1.4-2 Maximizing a weighted combination of the best k performers that participate

Whereas the contest specialization defines the solvers' objective, the seeker's objective is defined by how many solutions she ultimately needs. In practice, a seeker is typically interested in the upper tail of solvers' outputs as well as in generating multiple high performing submissions, also known as *candidate solutions*. Consider a typical innovation contest where the seeker is interested into a subset of top performing ideas out of which he would reward the top three (Terwiesch and Ulrich, 2009). Similarly, consider the job market hiring process of an academic institution that has opened a faculty position. A pool of applicants who self-select and compete in this contest face a non-negligible participation cost related to their job search. The Hiring Committee (the "seeker") wishes to attract a number of highly performing scholars and invite them for a fly-out ("candidate solutions"), out of whom the best one would be typically offered the faculty position (the "award").

As such, we allow the seeker to maximize the best k candidate solutions submitted, for an exogenously fixed number $k \in \{1, \ldots, N-1\}$. When k is small (e.g. an individual who wishes to develop a well-specified logo on 99designs.com), the seeker benefits more from developing a few "star performers" (e.g. outstanding logo submissions) than from marginal improvements in all of the submissions in a large pool of solvers. In contrast, a large k is indicative of a contest designer who does seek to improve many submissions.

Our next result shows that the optimality of the Winner-Takes-All (WTA) allocation is robust when the seeker optimizes the best or the top k *participating* candidate solutions.

Theorem 5 (Best kth participating outputs)**.** *Fix a number $k \in \{1, \ldots, N-1\}$ of candidate solutions. The Winner-Takes-All (WTA) allocation maximizes the expected performance of the top k participating solvers.*

Theorem 5 proves that the WTA is optimal when the seeker is interested in the top k participating candidate solutions. However, when $k = N$ the WTA may not always be optimal as Theorem 3 demonstrate. That is, for any value of $k \in \{1, \ldots, N\}$ we can summarize the implications of Theorem 3 and Theorem 5 for the structure of the optimal allocation as follows.

The optimal MW reward allocation has the following structural form:

$$\underbrace{\underbrace{\frac{R}{m^*}, \ldots, \frac{R}{m^*}}_{m^* \text{ positive awards}}, \; 0, \ldots, 0}_{\text{up to } m_0^* \text{ positive awards}}, \; \underbrace{0, \ldots, 0}_{\text{no awards}} \tag{1.6}$$

That is, the optimal allocation contains no more than m_0^* awards. As we demonstrate in §D.4, the optimal general reward allocation exhibits a similar structure and is not necessarily a MW allocation. In the special case where the seeker is interested in as many candidate solutions as solvers' total population (i.e. $k := N$) the setting would translate to a seeker who optimizes the total performance of the participating pool and we recover the objective already considered in §1.4-1. Further, WTA is optimal for all $k \leq N - 1$.

1.5 Conclusion

In this chapter, we study how innovation contest organizers can manipulate the number and size of rewards they offer to steer solver incentives for participation and effort. Our model extends existing theory and we offer a causal explanation for the prevalence of multiple awards in many crowdsourcing platforms in practice. For instance, in the ideation contest marketplace Tongal.com, seekers divide their budget equally among winners, and the number of offered rewards varies with a median of four, and a maximum of 50 rewards (see Table 2 of Kireyev, 2016). Further, the empirical analysis of Kireyev (2016) does not find evidence of risk-aversion among solvers which could be a reason to offer more than one reward based on Kalra and Shi (2001). Instead, Kireyev (2016) suggests the existence of incomplete information, or solver heterogeneity, as a possible interpretation for offering multiple rewards in Tongal.com. The latter is in line with our theory. Indeed, multiple awards are needed when solver participation is voluntary and performance is affected by both ability and effort, even when all parties are risk-neutral and cost of effort is linear.

On the managerial side, the effective budget allocation in an innovation contest requires a deeper understanding regarding three factors that confound decision making: the degree of contest specialization, the competitive nature of the contest and the objective of the manager. The interplay between a low degree to which solvers substitute ability for effort to enhance the performance of their solution, and an intense competitive landscape both make the incentive design problem challenging for the decision maker. However, the former decreases the value of offering various awards, whereas the latter requires more rewards of smaller size. Further, if the contest organizer cares about multiple candidate solutions the optimal budget allocation may be entirely different. The rationale behind these effects is of significant managerial value.

Our study of innovation contests coupled with methods that shed light on the strategic behavior of the solvers can form the basis for a more effective management of the innovation process. We view our work as an important step that can help academics and practitioners develop a better understanding of budget allocation at a strategic level.

This page is intentionally left blank

Part II.

Service Marketplaces

This page is intentionally left blank

Chapter 2

First Ranked First To Serve: Strategic Agents in a Service Contest[1]

Motivated by two-sided marketplaces and work-from-home contact centers that crowdsource demand to a pool of freelance agents, we model a service provider which ranks its agents in a predetermined number of priority classes based on their sales performance. The agents endogenously decide whether to participate and provide service on-demand. Agents' idle time is not compensated and better performers are utilized more and earn more. We study which priority structure maximizes profit in a markovian queueing model with random capacity. We show that a coarse partition with two priority classes is the optimal design of such a "service contest". Discarding available information on agents' relative rankings, or deploying coarser priority classes, provide better incentives for agents to participate, maximize firm's profit and asymptotically maximize welfare. This provides a game theoretic explanation for the extensive use of coarse priority rankings of freelance agents in work-from-home contact centers.

Key words: strategic servers; server priorities; work-from-home contact centers; service contest; service operations

[1]This Chapter is based on joint work with Serguei Netessine and Karan Girotra (Stouras et al., 2016).

2.1 Introduction

The recent rise of the sharing economy enabled service marketplaces to connect customers requesting high quality of service on-demand with geographically dispersed independent contractors. Operating similarly to ride-sharing platforms such as Uber or Lyft, *work-from-home contact centers* (firm) allow freelancers (agents or servers) to provide service on-demand while their idle time is not compensated. For example, working remotely at a work shift of his/her choice, a freelancer is able to handle calls, respond to emails, or comment to social media posts related to a client firm. Unfortunately, agent participation in such settings is not guaranteed as the agents, who have different skills (ability) as sales agents or service representatives, may find the cost to participate prohibitive. If and when a number of agents decide to participate (i.e. serve demand), the profit-maximizing firm would prefer the agents with the highest ability to enter, as well as to guarantee sufficient capacity to keep customers' wait low. In practice, to alleviate possible conflicts of interest that may arise when incentivizing both participation and high ability (as opposed to just participation), the firm offers the agents an incentive plan. Specifically, a work-from-home contact center allocates demand to agents *on priority* based on a predetermined number of priority classes to induce the agents to act in the firm's best interest.

To provide specific context for our setting, consider the case of LiveOps (Stouras et al., 2014), a work-from-home contact center that employs thousands of work-from-home independent contractors (i.e. operators who can not be called "LiveOps' employees" in legal terms) through their virtual marketplace. LiveOps acts as an intermediary between these agents and various organizations (clients) that outsource contact center services to them. Any incoming service demand for the client is then sourced to a pool of over 20,000 agents around the world. Depending on their individual work preferences, a subset of those agents will choose to work from home and serve calls for LiveOps' clients. However, unlike a traditional call center that routes an incoming call to *any* available operator chosen at random, LiveOps selects the *highest ranked available* agent, thus operating under a *service contest* business model.

A chosen priority ranking scheme makes a huge difference in work-from-home agents' earnings, as they are paid on-demand while their idle time is not compensated. For example, LiveOps could pay its agents \$13 for each hour of actively engaging with customers[2], and top performers may earn four times as much as poor achievers due to differences in utilization. Essentially, a work-from-home contact center transfers its idle time risk to its agents, who are willing to bear it in exchange for the flexibility it offers them to choose their own work schedule.

More broadly, many organizations employ the use of priority rankings in the form of *leaderboards* as a way to reward their agents and align their incentives with the firm. In the restaurant industry, waiters at the Massachusetts-based restaurant chain "Not Your Average Joe's" are ranked in terms of sales generated and better performers receive higher priority over incoming demand (Netessine and Yakubovich, 2012). As another example from retail, Percolata.com with 40 retail chains as clients, tracks sales per shop worker in its network and then ranks them and provides recommendations to the store manager on the optimal employee-mix to schedule at a given work shift to maximize profits (O'Connor, 2016). Similarly, online intermediaries

[2]See https://www.glassdoor.com/Hourly-Pay/LiveOps-Hourly-Pay-E105609.htm.

of the gig economy such as TaskRabbit, Lyft, Uber and Deliveroo leverage past rankings of their on-demand agents into their routing algorithms to better serve future demand (O'Connor, 2016).

A crucial mechanism that such service marketplaces employ to meet the needs of their clients is exactly the *incentive design* of the service contest. Our research questions are therefore centered around the following issues:

1. How do strategic agents behave in a service contest? How does allocating demand based on sales performance ranking impact their voluntary participation choices?

2. How should a work-from-home contact center optimally design a service contest? Specifically, what is the number and size of different priority classes to form, and how do they depend on the overall parameters of the service contest? Why do work-from-home contact centers rank their agents in a few priority classes in practice?

The purpose of this chapter is to study the optimal incentive design induced by server priorities to balance the participation and skill trade-off of the strategic servers. First, examine agents' equilibrium behavior for a chosen relative ranking scheme by the marketplace under a variant of the classical $M/M/N$ model with random capacity and congestion-sensitive demand. We extend this model accounting for priorities on the agents' side, and derive closed-form expressions for an associated performance metric, the expected utilization of each priority class taken over the random capacity available. We assume that the agents form beliefs on the anticipated participation actions of their peers and we show the existence of a unique symmetric self-confirming equilibrium (SCE; Fudenberg and Levine, 1993a) in which the agents can correctly predict the actual distribution of participants. Our analytical framework explicitly accounts for the unique aspects of a service contest such as (i) the endogenous participation choice of the agents whose sales performance capability is private information to them and is revealed to the firm ex post their participation decision, (ii) the amount of "contest rewards" (i.e. demand in the form of their utilization level) allocated to agents depends on the (endogenous) number of the agents that *actually* participate, and (iii) participating agents of higher type are weakly higher utilized. We find that agents' expected utility (strictly) increases in their sales performance capability in equilibrium. That is, more capable agents receive larger compensation in equilibrium. This implies that the way that agents self-select to participate in the service contest is characterized by a unique threshold that does not depend on the sales performance of each agent, and only agents that exceed that global threshold would find it rational to participate.

Second, we show that offering a coarse priority classes partition is beneficial for the firm. One may guess that in order to incentivize agents with heterogeneous ability to participate and generate the maximum profit for the firm, it may seem particularly unlikely to inject some coarseness into agents' priority classes partition. Allocating demand to participating agents in the most hierarchical way allows for agents with strictly higher expertise serve more demand, leading to a profit expansion. However, we show that a more refined partition induces fewer agents to participate compared to a coarse partition, so that any profit gains from more capable agents can not compensate for the decrease in congestion-sensitive demand due to insufficient capacity. Surprisingly, we prove that a very coarse, two-priority classes partition generates the highest profits for the firm. The latter insight offers a causal explanation related to incentives

alignment for why work-from-home contact centers in practice prefer to rank agents coarsely[3]. Also, we provide managerial guidelines on how should a work-from-home contact center adjust its system of priorities when the service contest parameters change.

Third, we show our results continue under various objectives of the firm. In particular, we show that two priority classes minimize the expected waiting time of the customers, and maximize the expected ability of the top performing agents. We also demonstrate that as the population of the agents grows large in an appropriate limiting regime, the number of participating agents converges in distribution to a Poisson distribution. Combining these results we prove that two priority classes asymptotically maximize system welfare.

2.2 Related literature

Our work combines the operations literature on user innovation and contests with recent papers on on-demand service platforms.

The operations literature has studied competitive settings called contests to drive innovation from agents who are not members of the firm (von Hippel, 2005). Building on the all-pay auction model of Moldovanu and Sela (2001), Terwiesch and Xu (2008) and Körpeoğlu and Cho (2017) show that separating the agents into two categories by offering one reward to the best performer and no awards to the rest (winner-takes-all, WTA scheme) maximizes the total effort of the agents, when everyone participates. Instead, the agents of our setting are competing for demand, are heterogeneous in ability and do not exert effort upon participating, i.e. they are in a service contest. Our work is also related to the literature on social comparisons with reference points (Roels and Su, 2013; Baron et al., 2015), since the compensation of an agent depends on the participation choices of his peers. However, our model does not capture the behavioral effects of feeling too far or too close from the top or experiencing loss aversion as Roels and Su (2013) and Baron et al. (2015) do respectively. We solve firm's problem of maximizing total profits of any *participating* agents by accounting for heterogeneity in agents' sales expertise while assuming that the population is large so that not everyone can participate. The novelty of our approach lies in treating the number of participants as a random variable, and hence the exact value of "rewards" in a service contest is only known *ex post* agents' participation in the form of realized utilizations of each priority class. That is, a work-from-home contact center re-distributes its total "budget" (i.e. available demand) in real time among the (endogenously) determined participants.

Further, in order to make a rational participation decision the agents form beliefs over their anticipated utilization level which we require to not contradict the observed outcomes in equilibrium. We use the self-confirming equilibrium notion developed by Fudenberg and Levine (1993a) that is a weaker form of the Rational Expectations equilibrium and is outcome-equivalent to Bayesian Nash equilibrium in our setting. See Su and Zhang (2008); Swinney (2014); Cachon et al. (2017) for applications of this equilibrium notion in supply chain management, inventory theory and marketplaces, respectively.

[3]Based on personal communication with LiveOps' executives, LiveOps offers several number of priority classes of agents based on their attributes called "pools" (cf. http://goo.gl/HQjrJH for some concrete examples).

Recently, on-demand service platforms have been modeled via a principal-agent framework in which the firm does not differentiate among the participating agents. In a multi-period setting, Gurvich et al. (2015) address capacity management questions with self-scheduling agents with exogenously fixed service rates, and Ibrahim (2015) uses appropriately scaled fluid limits to study the asymptotic behavior of such agents under various compensation plans offered by the platform. Cachon et al. (2017), Taylor (2016) and Tang et al. (2016) study the optimal dynamic pricing decision of a ride-sharing marketplace with ability-homogeneous agents and find the optimal surge pricing scheme to motivate sufficient agent participation in the presence of congestion. Gopalakrishnan et al. (2016) and Zhan and Ward (2015) analyze, respectively, routing and staffing decisions, and compensation schemes in many-server queuing models where servers can determine their service rates and servers' participation is controlled by the firm. Our work complements these papers in a single-period setting with self-scheduling capacity, in which the agents self-select to participate and a work-from-home contact center ranks its ability-heterogeneous agents into priority classes.

Most of the work in the queueing literature studies *customer* priorities and builds on Kleinrock (1967) who considers customers who arrive to an unobservable queue and bid to affect their priority assignment in the queue. While customer types are never pooled in this stream of research, Afèche and Pavlin (2016) show that coarsening the customer priority classes partition (i.e. pooling some customer types in a single priority class) optimizes revenue, and Hu and Benjaafar (2009) find that pooling available resources among different customers based on their service requirements is beneficial for customers' expected time in the system. Nazerzadeh and Randhawa (2015) show the asymptotic optimality of offering two customer priority classes to maximize revenue in large systems, a coarsening result that was recently extended and shown to further maximize social welfare (Gurvich et al., 2016). Focusing instead on the server-side analog, we find that offering two *server* priority classes maximizes firm's profit when servers' types are private information and their participation is voluntary. Notably, Knessl (2004) studies a storage allocation problem of parking spaces near a restaurant. By analyzing the $M/M/\infty$ queue with m primary servers and infinitely many servers of secondary priority, asymptotic results are provided by Knessl (2004) based on the earlier methods of Kosten (1937) and Newell (1984). In this chapter, we examine the steady state dynamics of the stable $M/M/\mathcal{N}$ queue with any number of (server) priority classes and $\mathcal{N}$ is a random variable with finite support representing the number of endogenously determined participating agents in equilibrium.

To the best of our knowledge, this is the first model of a work-from-home contact center which allocates work to agents on priority depending on their sales capability while agent participation is voluntary.

2.3 Model development

A work-from-home (virtual) contact center (*firm*, "she") serves incoming demand through a population of N independent contractors (*agents* or *servers*, "he") for a predetermined work shift (period) in order to maximize profits from sales of a given product. First, the firm decides on the number and size of priority classes (priority classes *partition*) to form by partitioning

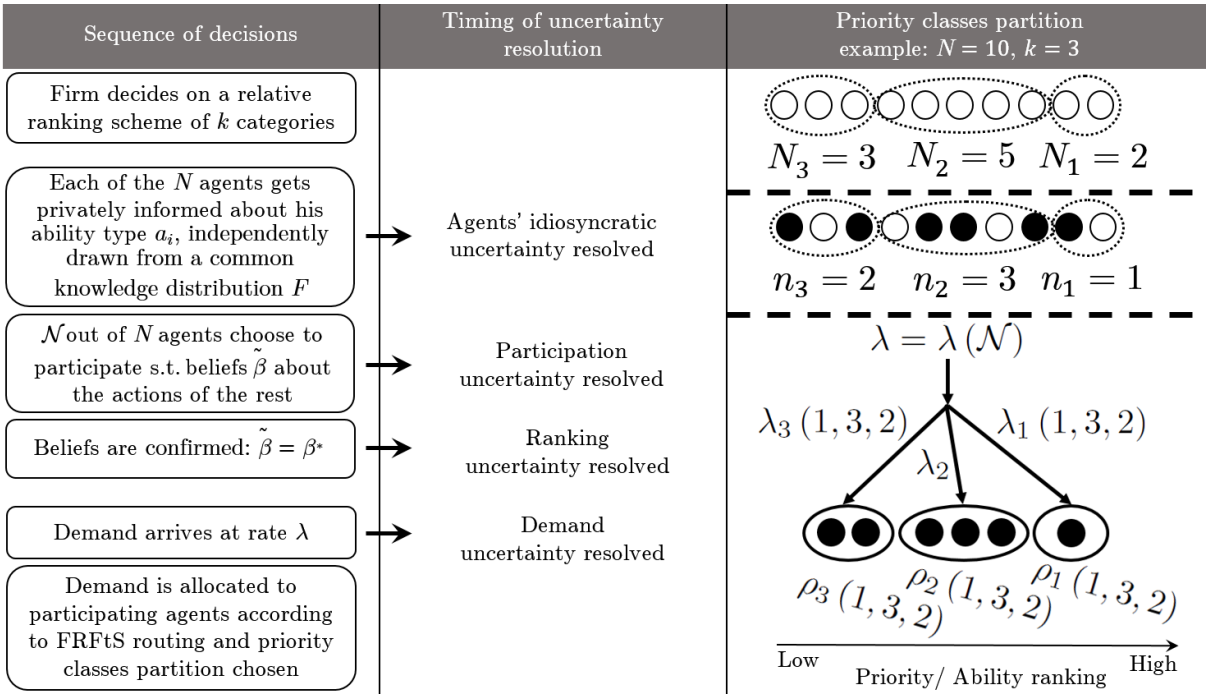


Figure 2.1 Sequence of decisions, timing of uncertainty resolution, and an illustration of the priority classes partition for a population of $N = 10$ agents split into $k = 3$ priority classes with $N_1 = 2$, $N_2 = 5$, $N_3 = 2$ agents respectively (Step 1). In this example, $n_1 = 1$, $n_2 = 3$, $n_3 = 2$ participating agents are realized for each class (Step 2). Incoming demand λ is shared among the participating agents according to priority classes formed and FRFtS routing (Step 3).

agents' population into k categories, where $k \in \{2, \ldots, N\}$. Second, the agents are privately informed about their sales or service capability (*ability*) and a subset of agents' population decides to work (participate) for the period. Participating agents are ranked by their ability according to the priority classes partition chosen and they are paid an exogenously fixed wage when utilized, while their idle time is not compensated. Third, each customer observes his expected waiting time of service as communicated by the firm, and decides whether or not to seek service upon experiencing a need for service. Due to its similarity with a "competition to provide service", we term our setting a *service contest*. We refer the reader to the right of Figure 2.1 for an illustration of the sequence of events and we describe below the steps of our model in greater detail.

We begin with the demand side. The customers experience a need for service according to a Poisson process with mean Λ, and are served by a realized number of $n \in \{0, 1, \ldots, N\}$ *participating* agents at exponentially distributed service times (see §E.1 for a list of the notation used). Consistent with anecdotal evidence from practice[4], work-from-home contact centers are typically supply-constrained, i.e. for a given mean demand for service Λ the number of potential agents satisfies $N < \Lambda$. Customers are processed on a first-come-first-served fashion, receive service value $V > 0$ and incur disutility $c > 0$ per unit of time waiting in the queue, and demand is sensitive to congestion. Specifically, upon experiencing a need for service and conditional that

[4]See https://goo.gl/ScetQI, assessed in February 2017.

$\{\mathcal{N} = n\}$ agents participate (in Step 2), a customer decides to seek service with probability q in order to maximize his expected utility of being served $U_c(q) := V - c \cdot W(\Lambda q, n)$ facing zero outside opportunity cost, where $W(\Lambda q, n)$ is the expected waiting time in queue of an $M/M/n$ system. After deciding to seek service, a customer does not abandon the system.

It is known (see e.g. Chapter 3 of Hassin and Haviv 2003) that conditional that $n \in \{1, \ldots, N\}$ agents have participated, the equilibrium demand arrival rate $\lambda(n) = \Lambda \cdot q(n)$ satisfies $\lambda(n) < n$ and is the unique solution $q^* < \frac{n}{\Lambda}$ to the equation

$$U_c(q) = V - c \cdot W(\Lambda q, n) = 0 \tag{2.1}$$

That is, observing the expected waiting time communicated by the firm based on the number of agents that choose to participate (in Step 2), delay-sensitive customers seek service (in Step 3) only at a stable rate (otherwise they experience negative utility). Operating similarly to a ride-sharing app that terminates service displaying "no drivers available at this point" to potential riders, a work-from-home provider announces service delays to customers. Conditional that no agents participate for the period, the customers do not seek service, i.e. $\lambda(0) = \lambda \cdot q(0) = 0$.

Next, we describe the supply side. The agents are heterogeneous in the probability they generate sales or create service value to the firm (*ability*) and follow an automated script. We normalize agents' service rate to unity reflecting the fact that the service duration is outside the control of the agents. The ability a_i of an agent $i \in \{1, \ldots, N\}$ is private information to him and it is not known by the firm when she decides on the relative ranking scheme to form in Step 1. Abilities are drawn independently of each other from the interval $[0, 1]$ according to a continuous distribution $F(\cdot)$ that is common knowledge[5] and strictly increasing in its support with density $f(\cdot)$. For example, some agents may be intrinsically more capable (thus having a higher a value) than others in resolving a specific technical request or persuading interested customers to subscribe to an insurance plan, or to buy a product.

In order to make a rational participation decision, agent i faces two kinds of uncertainties: he is unsure of the number of the agents who will choose to participate, as well as of his own rank-order among the participants. To model this "strategic uncertainty", we assume that the agents form (ex ante) *beliefs* about the anticipated participation actions of the other agents. Formally, agent i's belief $\tilde{\beta}_i$ is a probability distribution over the participation actions of the others, conditioned on his own participation action. In a *self-confirming equilibrium* (SCE; Fudenberg and Levine, 1993a) the agents can correctly predict the actual distribution of participating agents' actions β^*, that is $\tilde{\beta}_i = \beta^*$ in a SCE[6]. We assume *symmetric* (i.e. $\tilde{\beta}_i = \tilde{\beta}$ for each agent i) and *consistent* beliefs (i.e. the belief of each agent $i \neq j$, k is the same with the agent j's belief about k's actions), while excluding correlated beliefs. Hence, a self-confirming equilibrium in our setting is outcome equivalent to Nash equilibrium (Theorem 4 in Fudenberg and Levine, 1993a).

Based on a partition $\boldsymbol{N} := (N_1, \ldots, N_k)$ of agents' population decided by the firm in Step

[5] In practice, the firm has past sales data for each agent, as well as rankings on specialized training simulations for new hires. Such past data allow a work-from-home contact center to empirically estimate the distribution of sales or service performance across its agents.

[6] In what follows, we reserve the symbol $(\tilde{\cdot})$ to denote a belief, and denote an equilibrium action by a star $(\cdot^*)$.

1, the outcome of the service contest is affected by the individual participation choice of each agent, as well as by the choices of the others. Conditional that $n-1$ other agents participate let ρ_j be the realized utilization of agents in class j. We denote by w the hourly wage of the agents who are compensated on-demand for providing service[7], and by $c_p > 0$ agents participation cost into the service contest. Then, the utility of agent i is $w\,\rho_j$ if he participates and his ability a_i is ranked in priority class j, and c_p if he does not participate. In particular, agent i participates, if and only if, his *expected utility* from doing so

$$u\left(a_i;\,\tilde{\beta}\right) = \sum_{j=1}^{k} w\,\hat{\rho}_j\left(\tilde{\beta}\right)\cdot \mathbb{P}\left[a_i \text{ is ranked in priority class } j\right] \tag{2.2}$$

covers his participation cost c_p, where $\hat{\rho}_j\left(\tilde{\beta}\right) := \mathbb{E}_{\boldsymbol{\mathcal{N}}}\left[\rho_j\left(\boldsymbol{\mathcal{N}};\,\tilde{\beta}\right)\middle|\, a_i \text{ is ranked } j\text{th}\right]$ is the *expected utilization* of the agent ranked in priority class j, taken over the vector of participating agents in each class $\boldsymbol{\mathcal{N}} := (\mathcal{N}_1,\,\ldots,\,\mathcal{N}_k)$ subject to symmetric (ex ante) beliefs $\tilde{\beta}$ about the participating actions of the others. We note that Baron et al. (2015) consider a similar expected utility involving probabilistic beliefs over the actions of the rest agents to model a consumer's utility from consumption and the availability of the product due to the consumption choices of the other consumers.

We next describe the priority routing scheme we are focusing on. Motivated by what is often observed in practice in work-from-home call centers (Stouras et al., 2014), demand is allocated to participating agents based on the chosen partition $\boldsymbol{N}$ and according to a routing scheme that we term *First-Ranked-First-to-Serve* (FRFtS) routing. For any chosen partition, allocating demand under FRFtS routing induces a vector $\boldsymbol{\rho} := (\rho_1,\,\ldots,\,\rho_k)$ of realized utilizations for each class such that $\rho_j \geq \rho_{j+1}$ (the lower the index j, the higher the priority and utilization), while the n_j participating agents in class j are allocated demand rate λ_j, where $j = 1,\ldots,k$. We require the total demand allocated to match the incoming demand available, i.e. $\sum_{j=1}^{k} n_j\cdot\rho_j\left(\boldsymbol{n}\right) = \sum_{j=1}^{k}\lambda_j\left(\boldsymbol{n}\right) = \lambda$ for any realization $\boldsymbol{n} := (n_1,\ldots,n_k)$ of the number of participating agents in each priority class. By its very definition, FRFtS represents a form of efficiency compared to uniform routing because when several agents are free, better agents would be more utilized. In addition, when all agents are busy, a single queue is formed (pooled system) and customers experience the minimum wait (compared to a system with a dedicated queue in front of each server).

Finally, we describe firm's decisions made in Step 1 based on self-confirming beliefs on the anticipated actions of the agents. The firm is risk neutral and obtains an exogenously fixed revenue V_f when agent i is making a sale and pays him per-service wage w when he is utilized, as specified by the partition chosen. In Step 1, the probability that agent i is making a sale (i.e. his ability) is a random variable $\mathcal{A}_i$ for the firm. The firm determines the number and size of priority classes (*coarseness* level), or equivalently decides on k distinct and non-increasing values of expected utilizations $\hat{\rho}_1,\ldots,\hat{\rho}_k$ to form in order to maximize her *expected total profit*

[7]Note that the hourly wage w is exogenously specified and it is identical for all agents in our model who serve a given type of product, although in practice it varies by product (Stouras et al., 2014). Here, we focus on a simplified setting with one product only.

generated by any *participating* agents in equilibrium

$$\max_{k,\,(\hat{\rho}_j)_{j=1}^{k}} \Pi = \mathbb{E}\left[\sum_{i=1}^{N} \rho_i\left(\boldsymbol{\mathcal{N}};\,\beta^*\right)\cdot\left\{V_f\cdot\mathcal{A}_i - w\right\}\cdot \mathbb{1}_{\{\text{agent } i \text{ participates}\}}\left((\hat{\rho}_j)_{j=1}^{k};\,\beta^*\right)\right] \tag{2.3}$$

Note that the expectation operator in (2.3) is taken over any sources of randomness while only the participating agents generate profits for the firm due to the indicator function $\mathbb{1}_{\{\cdot\}}$. We term as *coarse* any partition with $k \in \{2, \ldots, N-1\}$ priority classes to distinguish it from the *fine* (or the most hierarchical) partition that has $k := N$ priority classes.

This chapter is organized as follows. We characterize work-from-home agents equilibrium behavior in §2.4. We then solve firm's problem in §2.5. Our basic model is extended in §2.6 in several directions. A table of notation used is provided in §E.1. We relegate all proofs and any technical results on the queueing dynamics of our setting in the Appendix and §E.2 respectively.

2.4 Agents' equilibrium behavior in a service contest

In this section we analyze the behavior of the agents in a self-confirming equilibrium focusing on symmetric pure strategies. Our first result, demonstrates that a work-from-home contact center cannot eliminate its priority classes entirely, i.e. a minimal degree of differentiation among the heterogeneous agents is required.

Lemma 3 (The value of priorities)**.** *Assume that the agents population size is sufficiently large so that:*

$$N > \lambda\left(N\right)\frac{w}{c_p} \tag{2.4}$$

If the firm decides to remove any priority classes, no agent has incentive to participate in equilibrium.

As we demonstrate in the proof of Lemma 3 if all agents are offered the same utilization regardless of their sales capability, either all agents or no-one participate in a symmetric pure equilibrium. Driven by the fact that work-from-home contact centers in practice source demand to a large pool of agents of the order of 20,000 (Stouras et al., 2014; LiveOps, 2014), we impose a condition that agent population is large compared to the maximum possible demand, the wage offered and agent participation cost. Alternatively viewed as $w < \frac{c_p\, N}{\lambda(N)}$ the condition (2.4) has also the economic interpretation that the wage paid is limited (or that the agents' population is large) so that it is not possible for *all* agents to participate, since they would not cover their participation cost (see also the Corollary 2 of Erat and Krishnan (2012) where a similar condition was shown to be sufficient to motivate N agents to participate in an innovation contest). How much of the agents' market that the decision-making firm chooses to incentivize to enter is then a non-trivial question. In what follows, we assume that (2.4) holds.

Lemma 3 implies that in order to guarantee non-zero participation and incentivize agents to enter the service contest, the firm should form *at least* two priority classes. Since work-from-home contact centers have typically a large agent base of the order of thousands, employing server priorities functions as a "toll" to control the participation incentives of selfish agents. Indeed, our next result shows that a priority classes partition acts as a screening mechanism

to filter the participants and prevent agents of sufficiently low ability from entering the service contest. This provides a game-theoretic reason to the firm for using server priorities, in addition to rewarding its top performers and punishing any low achievers.

Theorem 6 (Self-confirming equilibrium)**.** *Let* $\boldsymbol{N} = (N_1, \ldots, N_k)$ *denote the priority classes partition chosen by the firm in Step 1. There exist a unique* $p^* \in [0, 1]$ *and* $a_{min} \in [0, 1]$ *that solve*

$$\left.\begin{array}{c} \sum_{j=1}^{N} w\left(\hat{\rho}_j - \hat{\rho}_{j+1}\right) \cdot F_{N-j:N-1}\left(a_{min}\left(p^*\right)\right) = c_p \\ p^* = 1 - F\left(a_{min}\left(p^*\right)\right) \end{array}\right\} \tag{2.5}$$

such that only agents with ability $a \geq a_{min}$ *participate according to beliefs* p^* *on the fraction of participants, and we set* $\hat{\rho}_{N+1} := 0$*.*

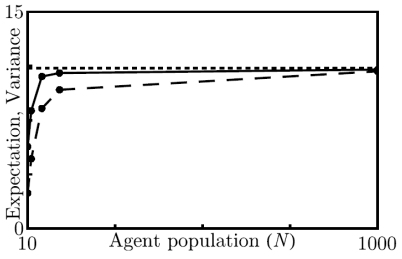
Alternatively viewed, the expression (2.5) is the participation (IR) constraint of the agents. Replacing the expected utilization of a given priority class with the *expected* (over the number of participants) value of awards of a contest gives an intuitive interpretation of the LHS of (2.5) using the language of contests and tournaments (see Moldovanu and Sela (2001), Terwiesch and Xu (2008) and Ales et al. (2017)). The expected compensation of the marginal agent in a service contest with (up to) k distinct levels of expected utilizations is equal to the expectation of the *differences* of expected utilizations across each possible rank he may achieve, subject to self-confirming beliefs.

Theorem 6 shows that for any priority classes partition determined by the firm that allocates demand according to FRFtS routing, the agents have a unique participation strategy in a SCE. Specifically, the agents enter the service contest, if and only if, their ability exceeds an ability threshold $a_{min} \in [0, 1]$ that does not depend on the realization of their privately known ability type, i.e. a_{min} is global. Following the sequence of events of our model, the agents know their own ability type when they choose whether to participate in Step 2 based on the announced priority classes partition in Step 1 (see Figure 2.1). Hence, by forming beliefs that turn out to be correct in equilibrium, the agents (and the firm) can accurately calculate the common ability threshold that uniquely characterizes agents' participation strategy. Theorem 6 is driven by the (weakly) monotone nature of FRFtS routing of demand to servers depending on their priority, namely that better ranked agents are weakly higher utilized and receive weakly higher expected earnings in order to cover their participation cost. Indeed, we show that agents' expected earnings over their potential rank-order is strictly increasing in ability. That is, once an agent ranked in a certain position finds it rational to participate, all higher ranked agents would participate as well. This implies the existence of a unique "marginal agent" with ability a_{min} who is indifferent between participating and his outside opportunity cost.

Agents' participation is central in our analysis and we investigate next the behavior of the distribution of the equilibrium number of participating agents. Due to the binary participation decision of the agents in symmetric strategies it is intuitive that the equilibrium number of participants follows the Binomial distribution with parameters the agents' population of the agents and their endogenously chosen probability to participate. Motivated by the fact that work-from-home contact centers employ a large-scale pool of independent contractors, we further characterize their asymptotic behavior by scaling the nominal arrival rate to $\Lambda \cdot N$ and agents'

(c_p, n_∞)	N	p_N	Maximum Deviation PDF	Maximum Deviation CDF
(0.6, 11.1)	10	0.56722	0.05883	0.07878
	20	0.40780	0.00703	0.07878
	50	0.21053	0.00071	0.00108
	100	0.10765	0.00023	0.00038
	1000	0.01101	0.00013	0.00018

(A)



(B)

Figure 2.2 (A) Convergence to a Poisson distribution. (B) Rate of convergence of the mean (solid line) and variance (trimmed line) of the participating agents as a function of agent population to a Poisson mean and variance $n_\infty = 11.1$ respectively (dashed line).

service rate to N. Note that such a scaling stochastically decreases the realized utilizations of each priority class.

Theorem 7 (Asymptotic behavior in equilibrium)**.** *(a) The number of participating agents in equilibrium $\mathcal{N}^*$ follows Binomial(N, p^*) with $p^* = 1 - F(a_{min})$. Further, agents' equilibrium participation probability $p^*(N)$ decreases in the size of the pool of agents N.*

(b) Set $\overline{m} := \max\left\{1, \left\lceil \frac{w\Lambda}{c_p} \right\rceil - 1\right\}$ and scale the nominal arrival rate to $\Lambda \cdot N$ and agents' service rate to N. As $N \to \infty$, the number of participating agents converges in distribution to Poisson(n_∞), where $n_\infty > 0$ is constant that does not depend on N and is the unique solution to the equation

$$\sum_{j=1}^{\overline{m}} w\hat{\rho}_j^{Poisson}(n_\infty) \cdot e^{-n_\infty} \frac{(n_\infty)^{j-1}}{(j-1)!} = c_p, \tag{2.6}$$

where $\hat{\rho}_j^{Poisson}$ denotes the expected utilization of the jth ranked agent when the number of participating agents follows Poisson(n_∞), for $j \geq 1$.

In lieu of Theorem 6, Theorem 7(a) shows that only a number of top ranked agents participate in equilibrium, who are distributed according to Binomial $(N, 1 - F(a_{min}))$. Agents' participation probability $p_N = 1 - F(a_{min}(N))$, which further determines the "marginal agent", tends to zero as the size of agents' population grows large. Intuitively, as agents population size grows without bound, the agents feel that their chances of recovering their participation cost are minimal due to fierce competition and instead prefer to drop out. What is of interest is the *rate* of convergence of agents' participation probability to zero. Interestingly, as $N \to \infty$ the expected number of participating agents $N \cdot p_N$ approaches an upper bound, a finite positive number n_∞ which is the unique solution to eq.2.6. Our result on strategic servers has some similarity to Lariviere and Van Mieghem's (2004) convergence of the number of strategically arriving customers over multiple periods to a discrete-time Poisson process. Theorem 7 shows that *service* agents competing for demand in a single period under a self-confirming equilibrium would generate a participation arrival pattern that approaches a Poisson distribution.

Theorem 7 is a limiting result, so we now consider *how fast* the number of participating agents converge to a Poisson distribution for participation cost $c_p = 0.6$, wage $w = 10$, delay-sensitive demand characterized from $(V, c, \Lambda) = (1, 1000, 7000)$ and a partition with two priority classes

with $N_1 = 5$ agents. Figure 2.2(A) reports the maximum absolute deviation in the probability density function (PDF) and the cumulative distribution function (CDF) for each iteration of the population of agents $N = 10, 20, 50, 100, 1000$. In Figure 2.2(B) we illustrate the speed of convergence of the expected number as well as the variance (which converges slower) of participating agents which follows the Bin (N, p_N) distribution) to the expectation and variance of the Poisson distribution with rate $n_\infty = 11.1$ respectively.

2.5 The benefits of coarse priorities

Having characterized agents' equilibrium behavior for any priority classes partition chosen by the firm in Step 1 (see Figure 2.1), we now solve firm's problem of the best way to rank the agents into priority classes in order to maximize profit. As shown in Theorem 6 any priority classes partition acts as a screening mechanism to incentivize the agents of a sufficiently high sales capability to voluntarily participate into the service contest, with a participation probability that is moderated by firm's choice of partition.

Firm's problem is written as

$$
\begin{aligned}
\max_{k,\,(\hat{\rho}_j)_{j=1}^{k}} \quad & \Pi = \mathbb{E}\left[\sum_{i=1}^{N} \rho_i\left(\boldsymbol{\mathcal{N}};\, p^*\right) \cdot \{V_f \cdot \mathcal{A}_i - w\} \cdot \mathbb{1}_{\{\text{agent } i \text{ participates}\}}\left((\hat{\rho}_j)_{j=1}^{k};\, p^*\right)\right] \\
\text{such that} \quad & \sum_{j=1}^{N} w\left(\hat{\rho}_j - \hat{\rho}_{j+1}\right) \cdot F_{N-j:N-1}\left(a_{min}\right) = c_p && \text{(IR)} \\
& \sum_{j=1}^{N} \rho_j\left(\mathcal{N}\right) = \lambda\left(\mathcal{N}\right) \quad \text{a.s.} && \text{(Budget constraint)} \\
& \hat{\rho}_j \geq \hat{\rho}_{j+1}, \quad j = 1, \ldots, N-1 && \text{(Monotonicity)} \\
& p^* = 1 - F\left(a_{min}\left(p^*\right)\right) && \text{(SCE)} \\
& \mathcal{N} \sim \text{Binomial}\left(N,\, p^*\right) && \text{(2.7)}
\end{aligned}
$$

Note that we require the incoming demand to be matched with the available participating capacity for every realization almost surely, allocating demand according to FRFtS routing. The threshold participation structure of the agents' equilibrium strategy implies that if an agent i with unknown for the firm ability $\mathcal{A}_i$ finds it rational to participate, all agents with higher abilities participate as well.

Congestion sensitive demand implies that increasing participation increases the expected utilization of each priority class. Further, the (IR) constraint characterizes a_{min} and it is linear in firm's decision variables. That is, the firm's problem of choosing a priority classes partition to maximize the expected total profit of any participating agents in equilibrium becomes a Linear Program (LP). Its solution is remarkably simple: it is optimal for the firm to coarsely allocate demand to her agents based on *two* priority classes, and deliberately ignore any differences in ability among agents ranked in the same class.

Theorem 8 (Two priority classes of servers are optimal)**.** *The optimal priority classes partition contains N_1^* primary and $N_2^* := N - N_1^*$ secondary priority agents given by the solution of the*

system

$$\left.\begin{array}{c} N_1^* = \arg\max_{1\le j\le N}\left\{\frac{F_{N-j:N-1}\left(a^*_{min}\right)}{j}\right\} \\ \hat{\rho}_1 \cdot F_{N-N_1^*:N-1}\left(a^*_{min}\right) + \hat{\rho}_2 \cdot \left\{1 - F_{N-N_1^*:N-1}\left(a^*_{min}\right)\right\} = \frac{c_p}{w} \end{array}\right\} \tag{2.8}$$

where $\hat{\rho}_1 = \hat{\rho}_1\left(N_1^*, N_2^*; \lambda\right)$ *and* $\hat{\rho}_2 = \hat{\rho}_2\left(N_1^*, N_2^*; \lambda\right)$.

The intuition behind the optimality of a very coarse, two-priority class partition in our setting is as follows. Consider the most fine-grained partition with N distinct priority levels, and coarsen it, for example at the top, creating a new "high" priority class composed of the two highest ranked agents. This will only increase the utilization (thus the earnings) of the second highest agent, while also affecting the top ranked agent in the opposite direction (due to the demand "budget" constraint) since lower priority classes remained unchanged. The induced ability threshold of the coarse partition cannot increase because the value of the newly created utilization difference at the top is higher (i.e. the utilization of the third ranked agent vs. the second ranked), while the probability of getting it (i.e. to be ranked at least second highest) is the same. That is, the expected payoff of the marginal participating agent will increase compared to the initial fine partition, and thus, more agents will be eager to enter the system. Increasing participation attracts higher incoming traffic and benefits both the firm and the agents. The threshold participation strategy of the servers guarantees that the latter will keep the same top-ranked agents motivated to participate while attracting some additional lower-ranked agents. The above informal arguments imply that the optimal partition contains no more than two priority classes and are made rigorous in Theorem 8 which is a *distribution-free* result. In addition, Lemma 3 demonstrates that completely ignoring server priorities (or forming one priority class by equally allocating demand to any participants) leads to no participation incentives for the agents. Hence, it is optimal for the firm to create precisely two priority classes.

Theorem 8 offers a causal explanation for why work-from-home contact centers in practice prefer to rank agents coarsely. In service industries where differences in agents' rankings are driven by differences in their ability, the firm is better off throwing away information. Surprisingly, the optimal service contest takes the form of a coarse ranking of agents using two letter grades: high ability agents ("A") or low ability agents ("B") . In particular, an optimal amount of "A" agents all share the same amount of demand, while any participating, lower ranked "B" agents are strictly less utilized. Note that our analysis ignores any organizational costs of offering multiple priority classes. The presence of such costs would only strengthen the arguments in favor of offering two priority classes.

The contribution of Theorem 8 lies in the intersection of service systems and contest theory. Conceptually, we view the above result as the "dual", server-side analog of the (asymptotic) optimality of two *customer* priority classes to maximize revenue in large systems (Nazerzadeh and Randhawa, 2015). Related to the theory of *effort*-based contests, the celebrated winner-takes-all (WTA) contest design of Moldovanu and Sela (2001) is essentially a coarseness result with two reward classes: one "primary" class with the top contestant and a "secondary" class with the rest contestants receiving no awards. In addition, viewing effort as agents' "bid" in an all-pay auction, Moldovanu et al. (2007) show that ranking competing agents into two categories would generate at least half of the revenue of the most discriminatory, fine ranking when the

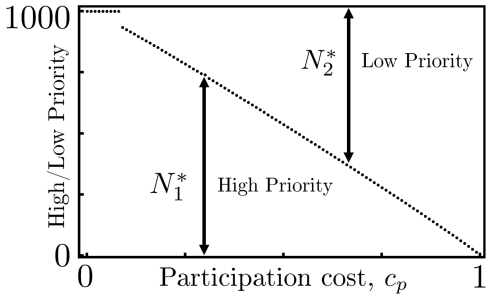


Figure 2.3 Dependence of the optimal number of top priority agents N_1^* (and low priority $N_2^* = N - N_1^*$) on the participation cost increase for Unif [0, 1] distribution of abilities and parameters $(N, w, \Lambda, V, c) = (100, 1, 200, 10, 2)$.

distribution of agents' abilities is "sufficiently convex". In that regard, Theorem 8 can be viewed as an extension of these results in an *ability*-based contest with endogenous awards (i.e. expected utilization of each class) determined by the strategic participation choices of the agents.

The optimal service contest contains two priority classes and depends on the exogenous parameters of the system, as well as on the expected utility of each individual agent. However, in large systems, as Theorem 7 demonstrates, the optimal allocation does not depend on the distribution of ability in the population.

For instance, in many relevant situations a work-from-home contact center may wish to induce its agents to work in a period with high cost to participate. Further, it is of interest to understand how should a manager of a work-from-home contact center react when more agents sign up into its platform increasing the agents' pool size. In particular, as agent's participation cost or agents' pool increase should the firm offer more, or fewer positions at the top priority class? The answer is not clear-cut and it depends on agents' incentives to participate.

Theorem 9. *The optimal number of top performing agents* (N_1^*) *is (weakly) decreasing in agents' participation cost* (c_p)*, and is (weakly) increasing in agents' population* (N)*. Asymptotically, the optimal allocation does not depend on agents' ability distribution.*

All else equal, faced with a higher incoming potential demand Λ, the expected number of participating agents of a work-from-home call center would flexibly increase. However, the dependence of the optimal number of primary priority agents (N_1^*) on an increase in agents' participation cost (c_p) is less straightforward. On the one hand, an increase in agents' participation cost implies that the firm would be able to attract fewer agents. By employing fewer (and thus more lucrative) high priority positions and reducing the *chances* of an agent belonging to one of them, the firm could choose to intensify the competition. On the other hand, the service contest designer could react by increasing the *number* of high priority positions, flattening the competition. Theorem 9 shows that the firm should optimally respond to an increase in agents' participation cost by offering the same or fewer top priority positions. In Figure 2.3 we illustrate the effect of the participation cost increase on the optimal number of primary and secondary priority agents. The reverse intuition holds for when the agents' population changes.

In addition, the optimal allocation is affected by the distribution of ability in the population, due to its dependence on the expected utility of the marginal agent. The incentive design of

a service contest for a product category in which talent is rare (e.g. selling specialized insurance products) is fundamentally different compared to a product category in which the agents compete with the "average Joe" (e.g. being a work-from-home call representative for Pizza Hut and competing for up-selling revenues). Somewhat unexpectedly, Theorem 9 establishes that asymptotically the optimal partition does *not* depend on the distribution of ability in the population. Specifically, for a sufficiently large population the behavior of the participating agents is characterized by agents participation cost and piece-rate compensation.

2.6 Extentions

In this section, we show the robustness of the previously obtained coarse priority classes partition that maximizes firm's profit under different objectives. First, we investigate a service quality objective of maximizing the expected value of an exogenously specified number of top sales agents among those who choose to participate, and an objective of minimizing the expected waiting time offered to customers.

Lemma 4. *Let N_1^* given by Theorem 8. Two priority classes with N_1^* primary priority agents:*
(a) minimize the expected waiting time of the customers.
(b) maximize the expected total ability of the top k participating agents, where $k \in \{1, \ldots, N\}$ is exogenously specified by the firm.

Lemma 4 shows that our previously obtained, extremely coarse priority classes partition is robust along the following two dimensions of service quality: expected wait in the queue and expected service quality offered due to the intrinsic capability of a service representative. Given that the agents make strategic participation choices, it is not clear a priori whether increasing agents' participation would alleviate customers' wait due to a negative externality these additional participating agents would impose on their peers who may in turn decide to drop out the system causing customers' wait to increase. Lemma 4 shows that this is not the case; in equilibrium, forming a partition that maximizes agents' participation increases the arriving traffic into the system while at the same increases agents' expected utilization which reduces customers' expected waiting time.

Surprisingly, two priority classes, i.e. the least informative ranking of agents by ability, maximize the expected ability of the top participating agents. One may expect that screening the agents in greater detail, or deploying a finer priority class, would increase the expected ability of the top participating agents. Again, our result is driven by the uncertainty the firm faces on the *number* of agents and their *ability* who will choose to participate. Increasing the number of agents who participate, or equivalently minimizing the marginal participating agent, strictly increases the support of the ability of the participants, and hence is beneficial for the work-from-home service provider. Lemma 4 is a distribution-free result because what matters only is the participating probability of the "marginal agent" which is the same across all distributions.

Finally, one may compare the optimal partition of the profit maximizing work-from-home contact center with the optimal partition that maximizes the total social welfare, i.e. the profit of the firm, agents' surplus net the expected waiting time of the customers. We show below

that our result is robust under this general objective, when the agents' population is sufficiently large.

Theorem 10 (*Two priority classes of servers asymptotically maximize welfare*)**.** *Let N_1^* given by Theorem 8. As agents' population grows large, two priority classes with N_1^* primary priority agents asymptotically maximize the welfare of the system, i.e. firm's profit and agents' surplus, net customers' expected waiting time.*

As agents' population grows large, their individual probability to participate tends to zero. In addition, the marginal participating agent, who merely covers his participation cost, tends to the top ability agent. Hence, for a sufficiently large population, only the top agent participates and covers his participation cost in equilibrium, i.e. earning zero expected utility. Further, from Theorem 8 and Lemma 4 we know that the optimal coarse partition maximizes firm's profit and minimizes the expected waiting time of the customers requesting service. Coupled with the fact that agents' payoff asymptotically approaches zero, we have that two priority classes of servers asymptotically maximize system's welfare.

Recently, a *customer* side analog of Theorem 10 has been discovered. In a limiting regime in which the arrival rate of customers and servers' service rate scale together, Gurvich et al. (2016) show that two customer priority classes maximize social welfare in large systems, extending the revenue-maximizing result of Nazerzadeh and Randhawa (2015). Although the loss of using two customer priorities is negligible in this limiting regime, Gurvich et al. (2016) show that there are important differences between the revenue maximizer and social planner in the level of "classification", i.e. the number of customers to allocate in each priority class in the asymptotically optimal partition. Instead, we show that, asymptotically, the work-from-home contact center and the customers extract the entire surplus, and hence the optimal priority classes partition of the profit maximizing firm is identical in our setting (in terms of "classification" and "coarseness") to the optimal partition of the social planner.

2.7 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 prioritize demand to their available capacity taking into consideration their strategic behavior and compensating them on-demand. Through a markovian queueing model with server priorities and random capacity, we analyze the steady state performance of a two-sided queue in equilibrium and evaluate the best way to rank the participating agents into priority classes. Our analytical results establish why various on-demand service platforms rank their independent contractors coarsely and strategically ignore available information on their agents' sales performance.

Chapter 3

Online Product Support Forums: Customers as Partners in the Service Delivery [1]

Organizations increasingly provide service to their customers via an online product support forum in which service delivery is partially delegated to an active community of users. Through an analytical model, we examine the relationship between impatient askers who post questions, types of questions the users choose to respond to and user response rate as a function of the service rate of the firm. Our results establish that it may be to the best interest of the firm to strategically reduce its service rate to boost a faster response rate from the community of users. We identify two key thresholds on asker impatience that suggest a different optimal service rate of the firm in the presence of the responses of the strategic users. This offers a game-theoretic explanation for why companies such as Microsoft and Apple with similar products and large online communities manage their online product support forums differently.

Key words: customer support; service contest; strategic servers; online communities; service operations

[1] This Chapter is based on solo-authored work; see Stouras (2016).

3.1 Introduction

Providing superior service is a challenging problem that often determines the sustainability of an organization. Firms have traditionally made substantial investments to maintain adequate capacity to serve demand and to regularly train their service representatives (Gans et al., 2003). However, the emergence of the internet allowed geographically dispersed users to collaborate and provide service through a global self-organized *online community* (Kraut et al., 2012). Thriving question-and-answer (Q&A) websites such as StackOverflow (Mamykina et al., 2011; Anderson et al., 2012) and Quora (Wang et al., 2013) let users post technical and general purpose questions respectively, and receive support by other users-members of the community. For instance, StackOverflow users can seek or provide help about programming software such as Python or Mathematica, and Quora members collaboratively create and share knowledge on healthy eating or business practices.

Recently, organizations with a large customer base such as Microsoft and Apple adopted *online product support forums*, as an innovative business model to serve customers by crowdsourcing service support to an active community of other customers, in addition to, the available service representatives of the firm. The key difference with the aforementioned third-party owned Q&A websites is that product support forums belong to the ecosystem of the respective firm, which is employing several agents to moderate its content. However, similarly to community-owned Q&A sites, Microsoft's Online Communities[2] and Apple Support Communities[3] partially utilize their large pool of users employing a contest-based incentive structure to provide fast and reliable service (Stouras et al., 2016; Terwiesch and Xu, 2008; Boudreau et al., 2011).

The focus of this chapter is on designing incentives for service in a product support forum that entails customers who abandon service, endogenous entry of strategic users as well as endogenous choice among multiple available "contests" that ran in parallel. In our model, the askers post easy and hard questions that have different awards and costs, and receive answers by the community of users as well as firm's servers. The askers are impatient, i.e. they abandon service after a random amount of time that may vary across easy and hard questions. The firm acts as a Stackelberg leader and first chooses her capacity correctly anticipating users' actions. The users follow by choosing their service rate as well as the probability to respond to each type of questions. Only answers received by the users during the random impatience time of a question and before the firm's servers resolve them are rewarded by the askers. The objective of the firm is to maximize askers' service value accounting for any associated staffing costs.

We aim to understand the optimal service strategy of a firm which partially crowdsources its service support to an active online community of users. Specifically, our research questions are centered around the following issues:

1. How do strategic community users determine their service rate and which questions to reply to, in the presence of firm servers that also provide service?

2. How should a firm manage crowdsourcing service to its online community? How often

[2]https://answers.microsoft.com/

[3]https://discussions.apple.com/

to participate to maximize the number of questions that receive an answer either by its servers or by the users of the online community?

3. How does askers' impatience affect users' and firm's optimal decisions?

We provide answers to these questions and make several contributions. We characterize users equilibrium service rate as a function of the service rate of the firm depending on the arrival of easy and hard questions and the strategic choice of the users among them. Our analysis demonstrates that the response rate of the users and firm's service rate initially behave as *complements*. However, for a sufficiently fast firm users and firm service rates become *substitutes* until a certain point where no user finds it rational to participate for any type of question.

Further, we show that despite any available high-cost-high-reward hard questions, the users mix their responses and often reply to low-cost-low-reward questions. We term such "mixed" equilibrium behavior as *exploration* to reflect the fact that the users respond to both types of questions with positive probability. For a sufficiently active firm the users' participation cost of resolving an easy question offsets any potential awards of reputation benefits for easy questions, and the users cluster their responses only under any high-cost-high-reward hard questions available. In that case we say that users perform *exploitation*, i.e. they respond only to one type of questions with the highest potential. An exploitation equilibrium outcome may be particularly inefficient when easy questions are swarming the system and outside users choose to resolve only the spare hard ones.

From the perspective of the forum manager, we show that there is always a unique service rate of the firm to maximize askers' service value net its staffing costs (Lemma 6). Interestingly, we find that askers' value is not always increasing with firm's service rate. This implies that it may be to the best interest of the firm to strategically reduce its service rate to boost a faster response rate from the community.

Motivated by the fact that online communities are typically large, we derive a closed form expression for firm's profit maximization problem. We prove that depending on askers' impatience level there are two key thresholds that essentially characterize firm's optimal capacity (Theorem 14 illustrated in Figure 3.7). For sufficiently low impatient askers, it is most beneficial for the firm to not resolve any posted questions and let the online community provide service. As askers' unwillingness to wait exceeds the first threshold, the firm gains from relying on users' support only to a limited extent and partially responding to questions with a two local maxima of capacity. The dominant service rate for the firm is determined by the cost of its staffing level contingent on the available traffic and users' explicit or implicit rewards. Finally, exceedingly impatient askers would discourage users from participating in which case providing service entirely in-house becomes necessary for the focal firm.

3.2 Literature

Our work combines research from recent papers in operations management studying service marketplaces and search among available alternatives with the theory of contests and all-pay auctions.

There are several papers in operations management that study strategic agents in services. Gopalakrishnan et al. (2016) and Zhan and Ward (2015) consider routing and staffing decisions of a service system in the presence of strategic servers who optimally balance a trade-off between their capacity cost and value of idleness. Accounting for customer abandonments and a large-scale self-scheduling workforce, Ibrahim (2015) characterizes the optimal staffing policy of the firm that sources work to a pool of on-demand servers. Also, in an endogenous participation contest model with incomplete information Stouras et al. (2016) characterize the optimal way to prioritize self-interested servers based on their performance when the system is stable with high probability. Further, Gans and Zhou (2007) assumes partial call center outsourcing, whereas Ren and Zhou (2008) study contracting issues when outsourcing calls to an outside service provider. None of these papers consider the outsourcing decision of a service firm to its customers, who rationally choose to act as servers and resolve firm's problems.

Our work is also related to the literature of search for the best alternative in a complex landscape. Weitzman (1979), in a seminal paper, modeled search as a sequential sampling process of independent alternatives and characterized the optimal policy seeking for the highest outcome. Employing a contest-based approach Erat and Krishnan (2012) examined the induced dynamics when a firm delegates the search for the best outcome to a pool of "outside" agents who endogenously choose among available contests upon entry. DiPalantino and Vojnovic (2009) study users equilibrium choice among multiple auctions via an all-pay simultaneous auction model, and Liu et al. (2014) extend these results conducting a randomized field experiment in a sequential all-pay auction model with complete information and exogenous participation.

Crowdsourcing contests are a powerful mechanism to boost engagement among agents to win an award to an announced competition. There is a vast literature in economics starting with Galton (1907) that has largely focused on what award structure offers the highest incentives for agents to exert effort accounting for potential information asymmetry among the contestants and the contest designer. Recently, Roels and Su (2013) study the optimal mechanism of incentivizing agents that are prone to social comparisons, while Terwiesch and Xu (2008) and Ales et al. (2017) examine the most efficient award structure to provide an innovative solution to a single posted and well-defined problem. In our setting, the users compete for service but they are capable of dynamically making endogenous participation decisions as well as strategic choices among the available alternative "contests" that run in parallel.

There is an increasing body of research in information systems and computer science that study community-owned Q&A sites. Driven by the abundance of available data from well designed and maintained Q&A sites such as Stack Overflow, empirical researchers analyzed users strategic behavior (Adamic et al., 2008; Anderson et al., 2012) in the presence of badges (Anderson et al., 2013) to promote valuable contributions from the community. Ghosh and Kleinberg (2013) consider a model of users' competition and endogenous participation in forums for education, while Jain et al. (2014) model sequential information aggregation for a single question to be answered while it is not costly for the users to contribute. We extend this literature in a dynamic model accounting for users' costly but endogenous participation and endogenous choice among possible questions of varying costs and benefits, conditional on entry.

3.3 Model

We consider a firm providing service over the finite horizon $[0, T]$ through an online product support forum composed by three distinct populations: customers who post questions (askers), a population of community subscribers (users) who are not affiliated with the firm and voluntarily provide service on-demand, and firm's employees (servers) who also respond to questions.

The *askers* post questions related to firm's products or services to the forum according to a Poisson process with rate $\lambda > 0$. We consider two kinds of questions: easy and hard, arriving at rates λ_e and λ_h respectively such that $\lambda_e + \lambda_h = \lambda$. Further, each asker is *impatient*, i.e. he abandons service if he does not receive an answer before a random amount of patience time that is IID across askers and questions' types following an Exponential distribution with mean $\theta > 0$. As shown by Baccelli et al. (1984) irrespective of the number of users or servers available, askers' abandonment makes the system stable.

There are $N \geq 2$ strategic *users* (or online community members) who are not affiliated with the firm but they are members of its online community. Each user i $(i = 1, \ldots, N)$ replies at exponentially distributed service times by simultaneously choosing (i) a service rate $\mu_i > 0$ to reply to questions, and (ii) a probability p_i to respond to easy questions. That is, user i's service rate to easy (resp. hard) questions is $\mu_i\, p_i$ (resp. $\mu_i\, (1 - p_i)$). Each time that a user responds to a question he incurs a cost c_e for easy (resp. c_h for hard) questions. Further, we conceptualize the users' decision to not participate by allowing the choice of $\mu := 0$.

There are various psychological, cultural or altruistic reasons that explain *why* users (i.e. firm's customers) provide service support to askers (i.e. other customers of the firm); see Jeppesen and Frederiksen (2006), Chesbrough (2013) and Kraut et al. (2012) for users' behavior in online communities, Nov (2007) and Yang and Lai (2010) in the context of Wikipedia contributions and Hamari et al. (2015) for the knowledge sharing economy. The askers posting questions in online communities such as the ones of Microsoft and Apple reward high contributors with reputation points that correspond to implicit prizes (e.g. the "Contributor of the Month badge", or "Level 9" user) and often explicit rewards (including product discounts and promotional coupons). In our model, we let v_e (resp. v_h) represent the total value of all these rewards to the users in terms of reputation points for answering an easy (resp. hard) question. Further, we assume that $1 < \frac{\lambda_e v_e}{c_e} < \frac{\lambda_h v_h}{c_h}$ so that the users have incentive to participate and derive higher relative benefit supplying an answer to a hard question compared to an easy one.

A user is rewarded by the asker of a given question with reputation points if and only if he provided an answer to the question before (i) the firm's servers respond, and (ii) before the asker decides to abandon service, whichever happens first. Indeed, in Microsoft's Online Communities all value-generating replies are shown under each posted question but if an asker decides to leave the system before an answer has been received, no later arriving answers are rewarded by the asker resulting in "unanswered" questions in which case Microsoft incurs a loss of goodwill cost. Similarly, once a Microsoft staff member responds to a question, any future responses by the users are redundant.

We denote by $\mathbb{1}_{A_{q\cdot}}$ the indicator function of the event $A_{q\cdot}$ of a user i $(i = 1, \ldots, N)$ being awarded for a given question of a given type (see Figure 3.1 for a graphical illustration). Let Q_e (resp. Q_h) be the set of the easy (resp. hard) questions posted by the askers over the finite

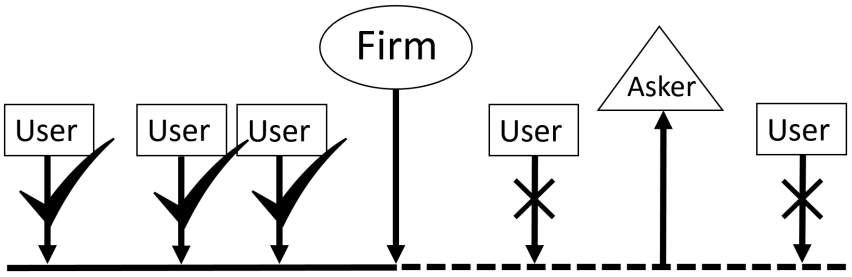


Figure 3.1 A product support forum model. All users that arrive before the firm and before the asker abandons service are rewarded by the asker; no other users are rewarded for the given question.

horizon $[0,T]$, and let $T_i(\cdot)$ be the set of times that user i responds to a question choosing a respective service rate. Then, user i chooses a rate μ_i and probability p_i (resp. $1-p_i$) to respond to easy (resp. hard) questions to maximize his expected utility given by

$$U_i(\mu_i,p_i)=\mathbb{E}\left[\sum_{q_e\in Q_e}\left\{v_e\mathbb{1}_{A_{q_e}(\mu_i,p_i)}\right\}-\sum_{t\in T_i(\mu_i\cdot p_i)}c_e+\sum_{q_h\in Q_h}\left\{v_h\mathbb{1}_{A_{q_h}(\mu_i,p_i)}\right\}-\sum_{t\in T_i(\mu_i\cdot(1-p_i))}c_h\right], \tag{3.1}$$

where the expectation operator $\mathbb{E}$ is taken over any sources of randomness. If user i decides to not participate into the forum (i.e. he chooses a rate $\mu_i := 0$), he receives a fixed utility normalized to zero. That is, in order for user i to find it rational to participate user i must attain $U_i \geq 0$.

Before the users make strategic decisions, the firm's employees (*servers*, or simply the firm) move first replying at exponentially distributed service times by choosing a service rate $s>0$ incurring a staffing cost[4] c_f per entry. Similarly to the users case, we allow the servers to choose a rate $s := 0$ to capture their non-participation decision. We note that the servers' rate s captures the firm's *total* capacity employing a workforce of an exogenously fixed number of servers each working at an identical rate s. The firm derives service value (or reputation benefits) V_e (resp. V_h) from each easy (resp. hard) question posted that receives an answer before its random impatience time. We assume that $c_f < V_e < V_h$. Finally, the servers are risk-neutral and choose a rate s in order to maximize the firm's expected utility of service which is the difference between the value generated from delivering satisfactory service to the askers and the cost for replying to questions:

$$\Pi(s)=\mathbb{E}\left[\sum_{q_e\in Q_e}V_e\mathbb{1}_{VC_{q_e}(s)}+\sum_{q_h\in Q_h}V_h\mathbb{1}_{VC_{q_h}(s)}-\sum_{t\in T_f(s)}c_f\right], \tag{3.2}$$

where $VC_{q_\cdot}(s)$ is the set of times that value has been created for each type of question, i.e. when a posted asker's question of a given type has received at least one answer during his patience time. We note that such an answer may arrive into the forum either by the N users who are not affiliated with the firm (i.e. at no cost to the firm), or by her servers incurring any associated

[4]That is, we are assuming that it is equally costly for the firm to provide an answer to an easy or hard question. This is a normalization for brevity of the exposition; our model can be extended to account for such a cost discrepancy.

staffing costs. Naturally, we assume that any answers received after askers abandon service are of no value to the asker resulting in poor service reputation for the firm.

We now summarize the sequence of events in the dynamic game played among the users and the firm. First, the askers post questions on the forum during the finite horizon $[0, T]$ and abandon service after a random time following Exponential distribution with rate θ. Easy (resp. hard) questions arrive at a rate λ_e (resp. λ_h). Second, firm's servers set a rate $s \geq 0$ (incurring associated costs) to serve demand, correctly anticipating users' behavior. Third, each of the N users simultaneously decide on their forum's participation rate $\mu \geq 0$ (incurring associated costs) along with a probability p (resp. $1 - p$) to respond to easy (resp. hard) questions. All answers posted before an asker abandons service or before the servers mark the question as resolved are rewarded reputation points accordingly.

3.4 Users' equilibrium behavior

In this section, focusing at a symmetric equilibrium we characterize users' entry rate and response pattern induced in the forum, and its dependence on the question type and servers' chosen entry rate. Our first result simplifies users' and servers' problems of visiting the forum over the whole interval $[0, T]$ into a per question decision.

Lemma 5. *At a symmetric pure equilibrium each user solves*

$$\max_{(\mu,\, p)\in[0,+\infty)\times[0,1]} \lambda_e\, v_e \frac{p\cdot\mu}{p\cdot\mu + s + \theta} - c_e\, p\cdot\mu + \lambda_h\, v_h \frac{(1-p)\cdot\mu}{(1-p)\cdot\mu + s + \theta} - c_h\ (1-p)\cdot\mu \quad (3.3)$$

At a symmetric equilibrium the per question expected utility of a user reflects the expected benefits of responding to easy (or hard) questions net his participation cost into the forum (eq. 3.3). The fractional terms indicate that the probability a user being rewarded is determined by whether the user chooses to respond to a given question, and whether his response was delivered before the question expires and before the servers arrive and mark the question as "resolved" (see Figure 3.1). Further, each time that a user responds to an easy (resp. hard) question he incurs a cost c_e (resp. c_h), whereas the chances of receiving a given award is affected by the probability he chooses a given type of question over the other one.

Following the sequence of events of §3.3, given a rate $s \geq 0$ set by firm's servers to participate into the forum each user simultaneously decides on a rate to participate together with the probability to respond to each type of questions. Theorem 11 characterizes the unique pure symmetric equilibrium of the service contest game played between the users and the firm.

Theorem 11 (Users' equilibrium behavior)**.** *(i) There is a unique pure symmetric equilibrium such that the users' rate to easy (resp. hard) questions is* $\mu_e^* := \left(\frac{1}{c_e}\sqrt{c_e\,\lambda_e\,v_e\ (s+\theta)} - (s+\theta)\right)^+$ *(resp.* $\mu_h^* := \left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\ (s+\theta)} - (s+\theta)\right)^+$*).*

That is, the users' (global) participation rate into the forum is

$$\mu^* = \mu_e^* + \mu_h^* \quad (3.4)$$

whereas if $\mu^* = 0$ *the users do not participate.*

Conditional on participation (i.e. if $\mu^* > 0$), the users' equilibrium probability of responding to easy questions is

$$p^* = \frac{\mu_e^*}{\mu_e^* + \mu_h^*} \tag{3.5}$$

(ii) The equilibrium number of user responses to easy (resp. hard) questions follows Binomial$(N,\, \mu_e^*)$ (resp. Binomial$(N,\, \mu_h^*)$).

Theorem 11(i) gives a closed form expression for the equilibrium behavior of the users for each type of questions arriving. Specifically, the rate μ_e^* (resp. μ_h^*) that the users respond to easy (resp. hard) questions is non-linear in firm's rate, and it has the same form for both type of questions. If users' best response is $\mu^* = 0$ then the users prefer to not participate. A positive μ^* rate indicates that the users decide to participate. Conditional on participation, the users choose to respond to easy questions with probability p^* given by (3.5), i.e. they resolve hard questions with probability $1 - p^*$. We note that if either one of the response rate to easy questions, or the rate to hard questions is zero it would indicate that the users resolve only a certain type of questions despite the possible abundance of questions of the other type into the forum. We explore the latter issue further in Proposition 3.

Further, a product support forum is a service system with random, on-demand capacity as in Stouras et al. (2016) and Ibrahim (2015). The forum users strategically decide on whether or not to participate into the forum, and conditional on participation they choose how frequently to participate. Each user arrives independently into the forum with rate μ^* and responds to easy questions with probability p^*. That is, we may think of users' responding to easy questions as performing N independent Bernoulli trials each having a "success" probability $\mu_e^* = \mu^* \cdot p^*$. Hence, the random number of users responding to each type of questions follows the Binomial distribution with the aforementioned parameters. We note that such a system is stable since the askers have positive probability of abandoning service (cf. Baccelli et al. (1984)).

The expressions of μ_e^* and μ_h^* in Theorem 11 imply that a higher cost of participation into the forum decreases users' equilibrium rate of responding to questions. Next, we investigate the less intuitive behavior of μ^* and p^* as a function of the firm's choice of interacting into her forum, as well as their dependence on exogenous parameters of the system.

Theorem 12 (Properties of μ^*)**.** *Let $s_{0,\,e} := \left(\frac{\lambda_e\, v_e}{c_e} - \theta\right)^+$, $s_{0,\,h} := \left(\frac{\lambda_h\, v_h}{c_h} - \theta\right)^+$. In the symmetric equilibrium, as a function of the rate s of the servers:*

(i) If $s_{0,\,h} > 0$ the users participate into the forum with rate $\mu^*(s) > 0$ given by Theorem 11 for each $s \in [0,\, s_{0,\,h})$, and do not participate for $s \geq s_{0,\,h}$. If $s_{0,\,h} = 0$ no user participates.

(ii) Let $s_e^* := \left(\frac{\lambda_e\, v_e}{4c_e} - \theta\right)^+$ and $s_h^* := \left(\frac{\lambda_h\, v_h}{4c_h} - \theta\right)^+$. The rate $\mu_e^*(s)$ (resp. $\mu_h^*(s)$) at which the users respond to easy (resp. hard) questions has a unique maximum at s_e^* (resp. s_h^*) and no user resolves easy (resp. hard) questions for servers' rate greater than $s_{0,\,e}$ (resp. $s_{0,\,h}$). Further, the global service rate of the users into the forum $\mu^*(s)$ attains a unique maximum $\frac{\left(\sqrt{\lambda_e\, v_e\, c_h} + \sqrt{\lambda_h\, v_h\, c_e}\right)^2}{8c_e c_h}$ at $s^* = \frac{\left(\sqrt{\lambda_e\, v_e\, c_h} + \sqrt{\lambda_h\, v_h\, c_e}\right)^2}{16c_e c_h} - \theta$.

(iii) Suppose that the askers are heterogeneous in terms of their abandonment rate for each type of questions, i.e. $\theta_e \neq \theta_h$. All else being equal, there exists a non-negative number s^* such that the global rate into the forum $\mu^*(s)$ has a unique maximum at $s^* = s^*(v_e,\, v_h,\, c_e,\, c_h,\, \theta_e,\, \theta_h)$ which is decreasing in c_e and c_h, and in θ_e and θ_h.

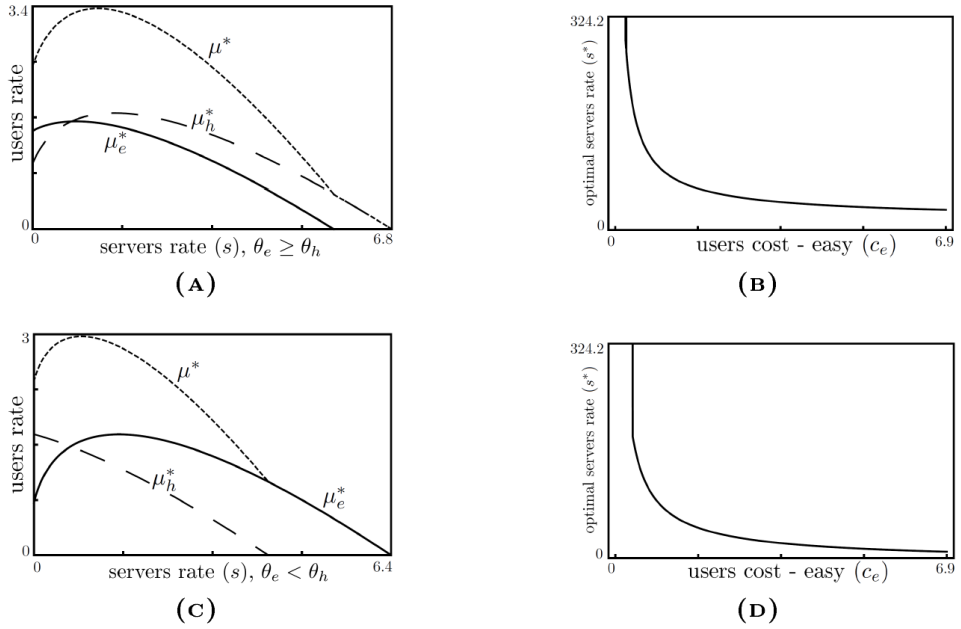


Figure 3.2 (A) and (C) Users' rate as a function of firm's rate; (B) and (D) Optimal firm's rate as a function of users' service rate cost for easy questions. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, c_e, c_h) = (13, 35, 2, 5)$, $(\theta_e, \theta_h) = (0.8, 0.2)$ for (A) and (B), and $(\theta_e, \theta_h) = (0.1, 2.8)$ for (C) and (D).

A number of insights can be inferred from Theorem 12 on the equilibrium structure of the game that users play competing for service against the firm. In equilibrium, no user responds to posted questions into the forum if the servers respond too frequently. Since participation is costly for the users, a very highly active server induces the users to drop out as they feel that their chances of winning an award are minimal.

The aforementioned firm's behavior holds also for each type of questions. In particular, there are non-negative rates $s_{0,e}$ and $s_{0,h}$ such that if the firm responds faster than $s_{0,e}$, no user resolves any easy questions and if they participate into the forum overall, the users only reply to hard questions. A similar effect holds for $s_{0,h}$, but no user participates for higher rate than $s_{0,h}$ (see also Figure 3.2(A) and (C)).

Theorem 12(i) shows a curious non-monotone effect of servers' activity level on the participation rate of the users in the forum. In particular, the equilibrium response rate of each user (μ^*) as a function of server's rate (s) is unimodal. As the servers increase their rate of responding from low to moderate values, initially the users' and server's rates behave as *complements* up to s^* that maximizes μ^*. Intuitively, in that initial phase the users are competing with the firm for awards from responding to questions. After s^*, a higher rate of resolving posted questions by the firm would slow down users' participation, i.e. μ^* and s become *substitutes*, until a certain value of servers' rate where μ^* becomes zero and the users are essentially giving up while only the firm resolves any posted questions.

The substitutability result of users-firm rates holds for users' global service rate into the forum (μ^*), and for their service rate at each type of questions as well (μ_e^* or μ_h^* respectively). Theorem 12(ii) proves that users' rate on easy or hard questions is unimodal as a function of the firm's rate. In Figure 3.2(A), (C) we illustrate the non-linear relation and substitutability of the

rates of the users and the firm for both the easy and the hard questions as well as for the global service rate of the users, when there are heterogeneous askers in terms of their abandonment rate for each type of questions, i.e. $\theta_e \neq \theta_h$.

In addition, we can explicitly solve for the optimal firm's rate that maximizes users' rate. As one can immediately infer from the closed form expression of s^* in Theorem 11(ii), firm's optimal capacity is strictly decreasing in users' service rate cost, and askers' impatience parameter. This result is robust even if we consider askers with varying unwillingness to wait depending on the type of question posted (Theorem 11(iii) illustrated in Figure 3.2(B) and (D)). The more costly it is for users to participate for a given category of questions, the slower the firm's capacity should be to maximize users' service rate.

So far we have investigated how firm's rate of responding to questions affects users' service rate into the forum. Conditional on participation, the users make strategic choices over the available easy or hard questions. Next, we show how firm's capacity can influence users' choice of questions resolved.

Proposition 3 (Properties of p^*)**.** *Assume that the askers abandon service with rate θ_e (resp. θ_h) when posting easy (resp. hard) questions, and that these rates are different. Let $m(\theta_e, \theta_h) = \min\left\{\left(\frac{\lambda_e v_e}{c_e} - \theta_e\right)^+, \left(\frac{\lambda_h v_h}{c_h} - \theta_h\right)^+\right\}$ and $M(\theta_e, \theta_h) = \max\left\{\left(\frac{\lambda_e v_e}{c_e} - \theta_e\right)^+, \left(\frac{\lambda_h v_h}{c_h} - \theta_h\right)^+\right\}$.*

(i) The users perform exploration, i.e. they respond to both types of questions with positive probability, when the rate of the servers satisfies $s \in [0, m(\theta_e, \theta_h))$. The users perform exploitation, i.e. they respond to one type of questions w.p. 1, when the rate of the servers satisfies $s \in [m(\theta_e, \theta_h), M(\theta_e, \theta_h))$. If $m(\theta_e, \theta_h) = 0$, the users always exploit for $s \in [0, m(\theta_e, \theta_h))$, while no user participates for $s \geq M(\theta_e, \theta_h)$.

(ii) Conditional on participation and irrespective of the impatient level of the askers, a higher cost of service for easy questions induces users to respond to hard questions with higher probability in equilibrium. The reverse holds as the cost for hard questions increases, all else being equal.

(iii) Suppose that it is equally costly for the users to respond to easy and hard questions (i.e. $c_e = c_h = c$) and that askers are equally impatient (i.e. $\theta_e = \theta_h = \theta$). Then, the equilibrium probability to respond to easy questions is strictly decreasing in firm's rate s for $s \in [0, s_{0,e})$.

Increasing firm's service rate initially motivates the users to respond to both type of questions with positive probability. Despite the presence of high-cost-high-reward hard questions, the users mix their equilibrium choices and often reply to low-cost-low-reward questions as well. We term such "mixed" equilibrium behavior as *exploration* to reflect the fact that the users respond to both types of questions with positive probability. Conceptually, users' equilibrium behavior into a forum resembles Erat and Krishnan (2012) result in innovation contests of searchers' clustering in specific regions of the solution space. Our results independently establish and offer a causal explanation for another form of clustering that arises in a dynamic service setting.

Proposition 3(i) identifies a potential equilibrium inefficiency stemming from users' strategic actions in an online forum. For a sufficiently highly active firm the users' participation cost of resolving an easy question offsets any potential awards of reputation benefits for easy questions, and the users cluster their responses only under any high-cost-high-reward hard questions available. In that case, we say that users perform *exploitation*, i.e. they respond only to

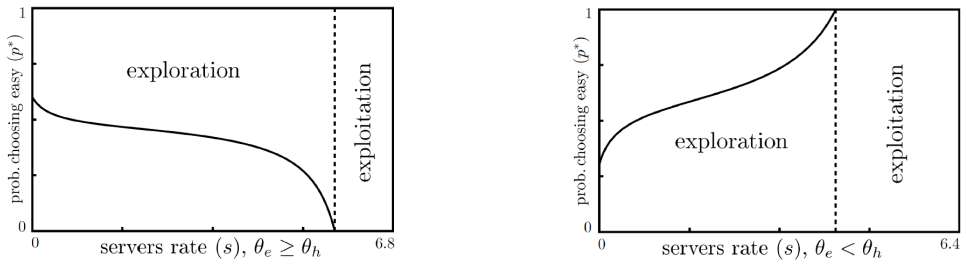


Figure 3.3 Users' equilibrium probability of replying to an easy question as a function of firm's rate. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, c_e, c_h) = (13, 35, 2, 5)$, $(\theta_e, \theta_h) = (0.8, 0.2)$ for (A), and $(\theta_e, \theta_h) = (0.1, 2.8)$ for (B).

hard questions. As shown in Figure 3.3 the users prefer to mix their responses in equilibrium among the available question types to maximize their gains. However, the self-interested users eventually choose the best outcome for them and respond to one type of questions w.p.1 when competing with a very active firm. In this scenario, the system may suffer from insufficient exploration and the firm cannot rely on outside users but instead has to commit costly internal resources to provide the desirable service support.

Intuitively, all else being equal, users should attempt the type of questions that offers them the highest upside potential (Gaba and Kalra, 1999). Although the users strategically choose the type of questions to respond among the available ones, there are various reasons for preferring to "exploit". For instance, the choice of users is affected by a potential low traffic for one type of questions, or by low reward-cost margins, or by a sufficiently low probability of being rewarded due to an actively participating server. From firm's perspective, an exploitation equilibrium outcome may be particularly inefficient when one type of questions are swarming the system and users of the firm's online community choose to resolve only the spare ones of the other type.

As users' service costs for easy questions rise higher, users increasingly abandon service for easy questions searching for hard questions with higher potential. In light of the unimodality result of Theorem 11, one might expect that for a sufficiently high impatient asker posting hard questions, as the cost of easy questions increases the users may initially prefer hard questions, but they may find it beneficial to switch to easy questions after a threshold. Proposition 3(ii) shows that this is not the case, and that with higher service rate cost of easy questions, more users choose hard questions that exhibit a higher potential. This effect is reversed as the cost of hard questions increases and more users are encouraged to attempt easy questions (compare Figure 3.4(A) and Figure 3.4(B)).

In addition, when it is equally costly to reply to each type of questions Proposition 3(iii) illustrates that as the firm's forum service rate increases, the users would still mix their responses among easy and hard questions while they will increasingly attempt more hard questions. As the firm resolves any available questions at a faster pace, easy questions of low-reward and low-cost become less attractive to the users who wish to cover their capacity and participation costs as well. From a managerial standpoint, resolving service requests too fast using internal resources may cause outsourced service support to self-interested agents to exploit the most beneficial outcome available while myopically ignoring attractive options of lower potential.

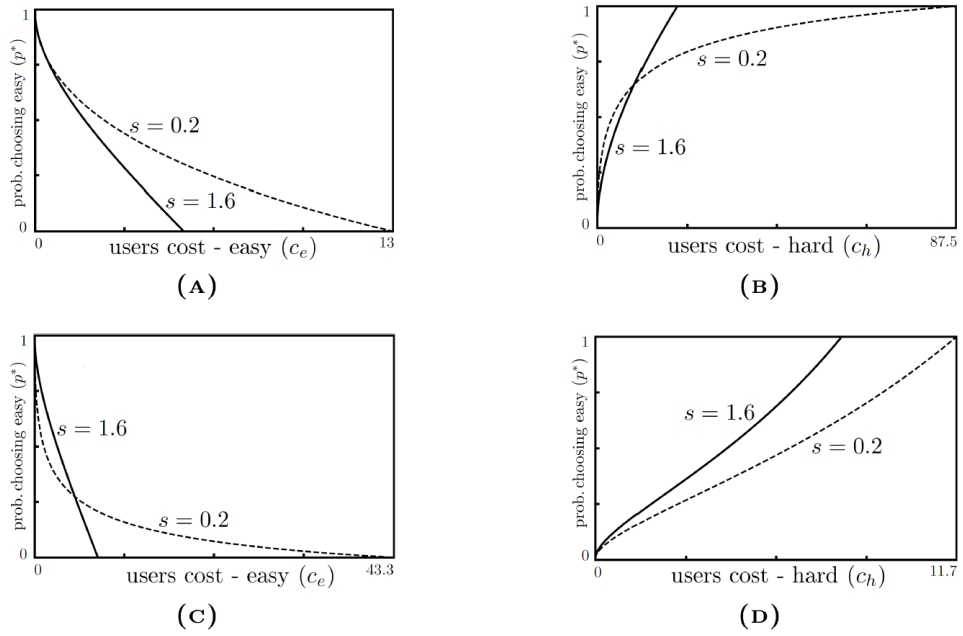


Figure 3.4 Users' equilibrium probability of replying to an easy question as a function of (A), (C) cost of easy questions, and (B), (D) cost of hard questions. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, c_e, c_h) = (13, 35, 2, 5)$, $(\theta_e, \theta_h) = (0.8, 0.2)$ for (A) and (B), and $(\theta_e, \theta_h) = (0.1, 2.8)$ for (C) and (D).

Proposition 4 (Clustering size)**.** *The number of user responses to hard questions is stochastically larger than the number of user responses to easy questions.*

Increasing the induced service rate in a platform with outsourced on-demand capacity may not always be desirable with self-interested users. As the firm increases her service rate there are two opposing forces at work: (a) initially users' rate increases competing with the firm for timely responses, and decreases otherwise, and (b) the users move away from easy questions, because a hard question becomes increasingly more attracting to the users. As Theorem 12 illustrates, the net effect leads to an initially higher rate of responding to hard questions at a decreasing rate of choosing an easy question. After that initial phase, a higher rate of firm's forum participation shrinks both users' participation and their choice of easy questions Theorem 12.

Due to the clustering effect caused by greedy forum users, more users will attempt hard questions over easy ones in expectation. Depending on the available traffic and related reward-cost variations this effect may be exacerbated causing little or no participation to one type of questions at the extreme.

We summarize the key characteristics of user equilibrium behavior in an online product support forum below before proceeding to analyze in §3.5 the most efficient management of an online forum:

(A) Users perform exploration of both type of questions for a sufficiently low active server.

(B) Exploitation of users' choices of hard questions emerges for a moderately active server.

(C) The service rate of a moderately active server and users' rate are compliments to the extent where a frequently operating server substitutes users' rate competing for service. No user participates in the presence of a rapidly responding server.

(D) A faster server causes users to choose hard questions more frequently, while she initially induces faster participation from the online community.

3.5 Managing a forum: How much to delegate to the community?

The results from the previous section highlighted the role of the firm's service rate into the forum to encourage users contribute to questions posted by the askers. In this section, we analyze how the firm should optimally manage her online forum. Specifically, we first investigate the optimal capacity decision for the firm in order to maximize her expected utility of service.

Lemma 6. *Let μ^* and p^* be the users' participation rate and probability of responding to easy questions in equilibrium. Then, the servers solve*

$$\max_{s\geq 0} \lambda_e V_e \frac{N\cdot(p^*\,\mu^*)+s}{(N\cdot(p^*\,\mu^*)+s)+\theta} + \lambda_h V_h \frac{N\cdot(1-p^*)\,\mu^*+s}{(N\cdot(1-p^*)\,\mu^*+s)+\theta} - c_f\, s \tag{3.6}$$

At a symmetric equilibrium firm's objective (3.6) reflects the net benefits of resolving an asker's service request in time (either by her servers or the users of the online community) and the cost for her servers visiting into the forum. Specifically, any of the N users of the online community or the servers who respond to a given question result in a happy customer, and subsequently create service value for the firm. However, tapping into an actively participating online community of users makes the firm realize these benefits at no staffing cost to its service operations.

Theorem 13. *Let $\mathcal{I}_1 := [0,\, s_{0,e}]$, $\mathcal{I}_2 := [s_{0,e},\, s_{0,h}]$ and $\mathcal{I}_3 := [s_{0,h},\, +\infty)$ and define*

$$R(s) = \begin{cases} \lambda_e V_e \frac{N\cdot\left(\frac{1}{c_e}\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta)}-(s+\theta)\right)+s}{\left(N\cdot\left(\frac{1}{c_e}\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta)}-(s+\theta)\right)+s\right)+\theta} + \lambda_h V_h \frac{N\cdot\left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)}-(s+\theta)\right)+s}{\left(N\cdot\left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)}-(s+\theta)\right)+s\right)+\theta}, & s\in\mathcal{I}_1 \\ \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{N\cdot\left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)}-(s+\theta)\right)+s}{\left(N\cdot\left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)}-(s+\theta)\right)+s\right)+\theta}, & s\in\mathcal{I}_2 \\ \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{s}{s+\theta}, & s\in\mathcal{I}_3 \end{cases}$$

Then, the firm participates into the forum at a unique rate s^*, where $s^* = \underset{s\in\mathcal{I}_1\cup\mathcal{I}_2\cup\mathcal{I}_3}{\arg\max}\,\{R(s)-c_f\,s\}$.

Theorem 13 outlines the optimization problem associated with firm's optimal choice of resources devoted to providing service in her online product support forum. The functions $R(s)$ and $c_f\,s$ are the "revenues" (i.e. askers' service value created) and costs when the firm resolves questions at a rate s. Although one may intuitively expect the askers' benefit to be increasing in firm's service rate (i.e., a faster responding firm will increase the overall rate that a given question is solved), the following numerical examples (shown in Figure 3.5) demonstrate that increasing firm's service rate can lead to *lower* service value generated for the askers when some users of the online community participate in the forum. This tension arises only in the regions $\mathcal{I}_1$ and $\mathcal{I}_2$ where both users and firm participate. Trivially, firm's revenue is increasing in s in region $\mathcal{I}_3$.

The intuition behind this non-monotone behavior of firm's revenue is as follows. Consider the case where users participate in an "exploration" phase, i.e. they solve both easy and hard questions with some positive probability. As the firm resolves questions at a faster pace, the

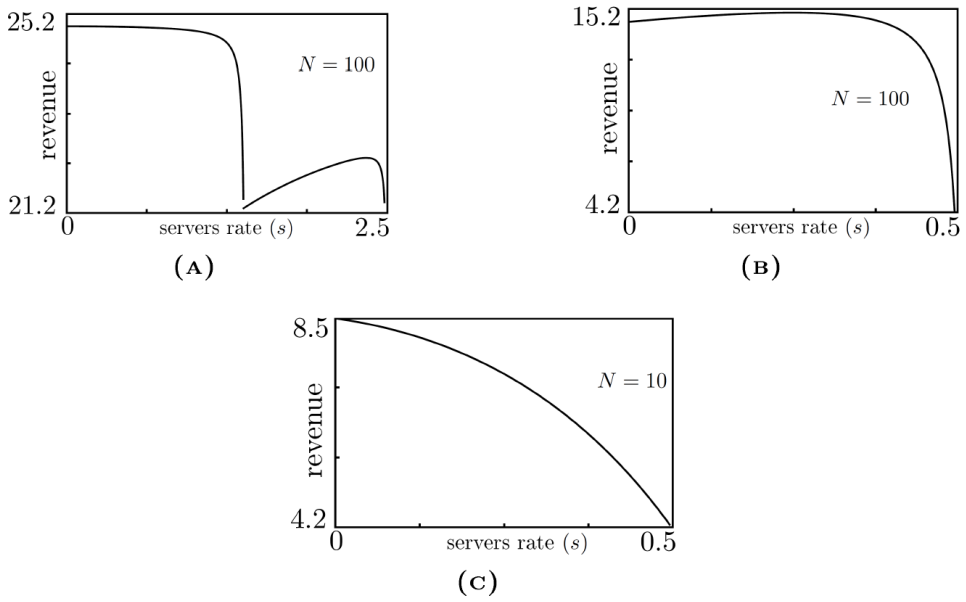


Figure 3.5 Firm's revenue as a function of servers' rate. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, V_e, V_h) = (13, 35, 10, 15)$ and $(c_e, c_h, c_f) = (6, 13.2, 2.2)$. Askers' impatience and users' population varies as follows: (A) $N = 100$ and $(\theta_e, \theta_h) = (0.8, 0.2)$; (B) $N = 100$ and $\theta_e = \theta_h = 2.2$; (C) $N = 10$ and $\theta_e = \theta_h = 2.2$.

users initially increase their service rate behaving as substitutes to firm's rate increase. In that region the firm's revenue naturally increases with firm's service rate. Then, increasing firm's capacity will decrease users' rate and at some point this decrease more than offsets the increase in askers' benefit from a faster firm, hence asker's benefit (and firm's revenue) drops. Hence, even if labor is cheap and staffing costs are negligible, the firm prefers to not interact or to resolve questions very slowly! This effect is even more pronounced when the online community is small (Figure 3.5c).

We state the above observations as the following proposition.

Proposition 5. *In the presence of a participating online community, firm's revenue is non-monotone in firm's capacity.*

Theorem 13 shows that firm's problem has a unique solution for all feasible values of the exogenous parameters of the system. Motivated by the fact that online communities are large[5] in practice, we describe the optimal strategy for the firm in closed-form as follows.

Theorem 14. *Assume that there is a sufficiently large online community of users. Then, depending on askers' impatience level, the firm's utility is maximized at*

$$\Pi^* = \begin{cases} \lambda_e V_e + \lambda_h V_h, & 0 < \theta < \frac{\lambda_e v_e}{c_e} \\ \max\left\{\lambda_e V_e + \lambda_h V_h + c_f \theta - 2\sqrt{c_f \theta \lambda_e V_e},\quad \lambda_e V_e + \lambda_h V_h + c_f \theta - c_f \frac{\lambda_e V_e \theta}{\lambda_h v_h} - \lambda_h v_h\right\}, & \frac{\lambda_e v_e}{c_e} \le \theta < \frac{\lambda_h v_h}{c_h} \\ \lambda_e V_e + \lambda_h V_h + c_f \theta - 2\sqrt{c_f \theta\,(\lambda_e V_e + \lambda_h V_h)}, & \theta \ge \frac{\lambda_h v_h}{c_h} \end{cases}$$

attained at $s_1^* = 0$, $s_{21}^* = \left(\sqrt{\frac{\lambda_e V_e \theta}{c_f}} - \theta\right)^+$ *or* $s_{22}^* = \left(\frac{\lambda_h v_h}{c_h} - \theta\right)^+$, *and* $s_3^* = \left(\sqrt{\frac{\theta\,(\lambda_e V_e + \lambda_h V_h)}{c_f}} - \theta\right)^+$, *respectively.*

[5]For instance, Microsoft Online Communities have more than 2,000 active users who voluntarily provide service in their free time to other customers (see `https://goo.gl/OQkkFH`, accessed on September 23, 2016).

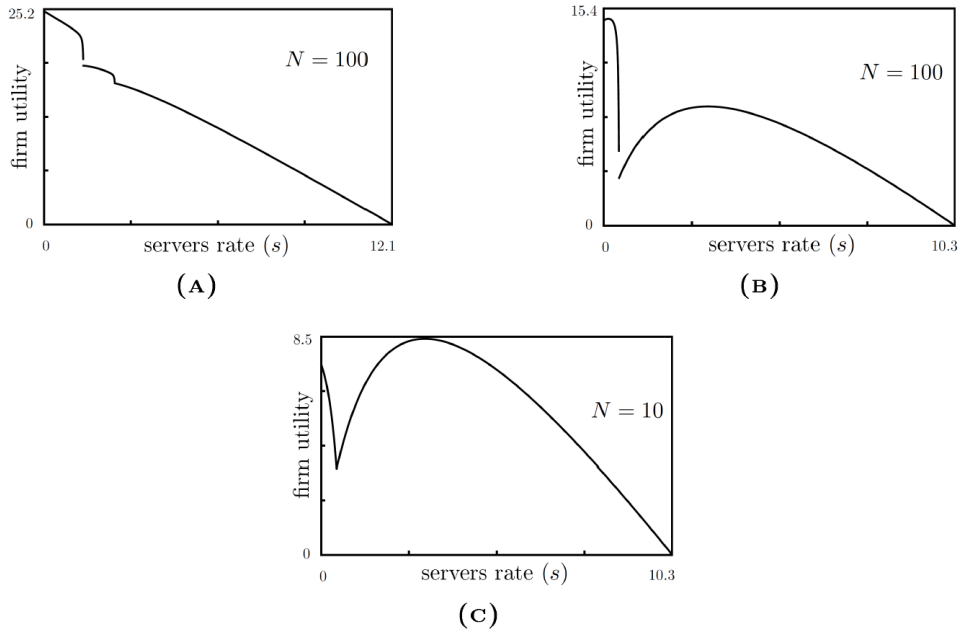


Figure 3.6 Firm's utility as a function of servers' rate. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, V_e, V_h) = (13, 35, 10, 15)$ and $(c_e, c_h, c_f) = (6, 13.2, 2.2)$. Askers' impatience and users' population varies as follows: (A) $N = 100$ and $(\theta_e, \theta_h) = (0.8, 0.2)$; (B) $N = 100$ and $\theta_e = \theta_h = 2.2$; (C) $N = 10$ and $\theta_e = \theta_h = 2.2$.

Theorem 14 characterizes the optimal strategy for the firm when it has the option to outsource service to an abundant online community of users available. In particular, there are two thresholds in askers' impatience level that provide a simple rule of thumb for the desirable service outsourcing level. For sufficiently low impatient askers, it is most beneficial for the firm to not resolve any posted questions and let the online community provide service. As askers' unwillingness to wait exceeds the first threshold, the firm gains from relying on users' support only to a limited extent and partially responding to questions with a two local maxima of capacity. The dominant service rate for the firm is determined by the cost of its staffing level contingent on the available traffic and users' explicit or implicit rewards. Finally, exceedingly impatient askers would discourage users from participating and then providing service in-house becomes advantageous for the firm. Refer to Figure 3.7 for a graphical illustration of the optimal strategy of the firm.

Online product support forums typically have a large number of users who interact with themselves and the contents of the website, generating as well consuming online content. In Figure 3.6(a) we plot firm's objective as a function of its rate assuming that there is a medium size online community ($N = 100$ users) with a heterogeneously impatient population of askers. For the given parameters of the system it is optimal for the firm to not participate and let the users (i.e. its customers) to respond to its customers. However, for less willing to wait askers some interaction by the firm is needed to motivate its users to participate frequently. This is in sharp contrast with the case of a small online community ($N = 10$) where in order to offer superior service to her askers the firm has to maintain a large capacity that essentially makes the few users of the community to drop out of the system (compare Figure 3.6(b) with Figure 3.6(c)).

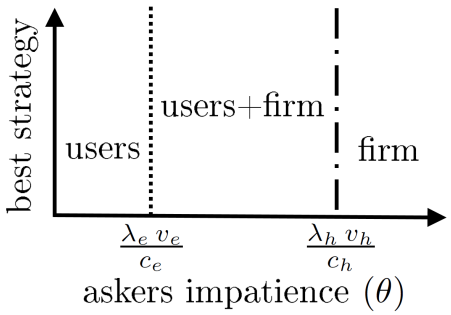


Figure 3.7 The optimal management of an online product support forum.

Theorem 14 also shows that demand for service of either the easy (λ_e) or hard questions (λ_h), and firm's staffing cost c_f have an intuitive effect on the optimal capacity level, when it is optimal for the firm to interact into the forum and partially delegate service to its active online community. Specifically, as the traffic increases or as the staffing cost c_f decreases, the firm benefits by increasing its service rate.

We summarize the managerial implications of Theorem 14 on whether it is optimal for a firm to crowdsource its service operations in Figure 3.7. If the firm is best served by either resolving all questions internally or totally outsource service to its online community (left or right region in Figure 3.7, respectively), the effect of askers' impatience reflects what discussed in §3.4. In particular, the asker's abandonment rate on users' and firm's service rates acts similarly to an additional participating firm. For a sufficiently patient asker, abandoning service faster will initially induce users to boost their service rate for both easy and hard questions, up to a level where the rate of responding in at least one of these question types will drop until it is no longer beneficial for a user to respond at all. Similarly, confronted with highly demanding askers the service provider should design its reputation-based incentive scheme appropriately so that to attract responses from the abundant users of the online community in order to alleviate the service congestion.

3.6 Conclusion

Motivated by the growing business model of organizations outsourcing service and product support to an active online community of users to provide fast service to their customers, we develop a formal analytical model that helps understand how an online forum should be managed. Our modeling framework captures several unique aspects of such an innovative model for service delivery, including (i) askers' unwillingness to wait, (ii) the potential incentive misalignment between firm and interacting users, and (iii) the extent to which firm's participation affects users' choice of questions available and rate of participation. We identify the following key characteristics of using customers for customer service summarized below:

- *Exploration-exploitation.* Users perform exploration of both type of questions for a sufficiently low active server, while users exploit and respond only to one type of questions for a moderately active server.

- *Substitutability of users-firm rates and endogenous participation.* The service rate of a

moderately active server and users' rate are compliments to the extent where a frequently operating server substitutes users' rate competing for service. No user participates in the presence of a rapidly responding server.

From a managerial perspective, our results summarized in Figure 3.7 provide a simple rule of thumb for when outsourcing service delivery to customers is desirable for the firm or even optimal. For sufficiently high willing-to-wait askers, it is most beneficial for the firm to not resolve any posted questions and let the online community provide service. However, for moderately impatience askers the firm gains from relying on users' support only to a limited extent and partially responding to questions. Further, when coping with exceedingly impatient askers providing service entirely in-house becomes the most advantageous strategy for the firm.

We believe that the aforementioned insight offers a causal explanation on why Fortune 100 companies such as Microsoft or Apple interact differently into their respective product support forums depending on the type of question being asked. Future research should test the validity of our model and its assumptions on empirical grounds. It would be interesting to see if our results continue to remain valid when users have a varying skill level.

In this chapter, we studied a simultaneous move game of endogenous participation of users with endogenous choice of available alternatives. Further, the forum users are assumed homogeneous in terms of their expertise which is a simplification of reality. Our work models the first order effect observed in a company's product support forum where all qualifying answers accumulate reputation points over time from future potential askers who find the answers useful. In many cases, the forum users have equal chance to get rewarded since they are often highly uncertain of the asker's subjective evaluation of their response. We leave to future research to examine the strategic user behavior considering a sequential model of endogenous participation and choice, and heterogeneous users with privately known skill levels (see Jain et al. (2014) and Liu et al. (2014) towards that direction).

A further direction is introducing a sequential model with learning dynamics that combines strategic askers with self-interested users with endogenous entry and endogenous choice among multiple postings. Albeit challenging, such an approach could describe askers' strategic stopping decision of the successive arrivals of answers to a posed question. Terminating the incoming answers too fast may resolve asker's question, although at a potentially lower quality compared to a belated response from a top-rated user.

This page is intentionally left blank

Conclusion

Contests offer a promising alternative to traditional incentive schemes for resource allocation in distributed marketplaces with self-interested agents who participate voluntarily. However, the voluntary participation nature of the independent agents of these systems necessitates a re-thinking of traditional contest theory because direct extensions of incentive structures that work well in small and fixed populations can fail in large-scale on-demand marketplaces. My thesis is that it is necessary to take an explicit "on-demand" approach to contest design. Indeed, the value of contest design theory in on-demand marketplaces will depend on its efficacy to balance the incentives of independent agents with the objectives of the marketplace, as well as to create value for the overall ecosystem that the marketplace operates. Once the incentives are successfully aligned, contests may provide an efficient and easy-to-implement solution to a variety of problems that on-demand marketplaces face.

To date, most of the literature on on-demand marketplaces has focused on incentives to attract sufficiently *many* participants ("quantity" of participants) through the design of the optimal pricing scheme. Indeed, current ride-sharing business models offer a surge pricing mechanism to motivate drivers to explore geographic locations with increased demand for a ride. In this dissertation, I argue that the design of *skill* incentives should be incorporated into the incentive design to maintain a high degree of service quality ("quality" of participants). Further, agents of a broad class of on-demand marketplaces need to be incentivized to take a certain *action*, conditional on participation ("effort" of participants). On-demand marketplace designers should incorporate all these three strategic effects to improve the efficiency of their platforms.

I adopt a *process view* of the decisions of the agents and the marketplace illustrated in Figure 3.8. This involves attracting a sufficiently large, heterogeneous agent population, motivate high participation and attract high performers, incentivize high output among the participants and focus on the key subset of the participants (target group) to optimize a given objective. In contrast, a traditional organization who has full control over its workforce is typically restricted to staffing ("how many employees to hire?") and pricing ("how much to pay them?") decisions. Given that staffing decisions are placed well before the demand is realized, the skills of the employees hired for a specific task exhibit a much higher degree of homogeneity compared to the independent contractors (agents) of a marketplace who are utilized *after* the demand realization.

In this dissertation, I develop a theoretical framework that describes the strategic behavior of agents and the optimal design of agent incentives in on-demand marketplaces to optimize the objectives of the marketplace and the social planner. A conceptual contribution of my work is that to attract the *right* participation and output, a marketplace should discriminate among its agents. In a meritocracy-based organization, higher performers are higher utilized and rewarded, while low achievers are eventually screened out. The latter can be seen as a

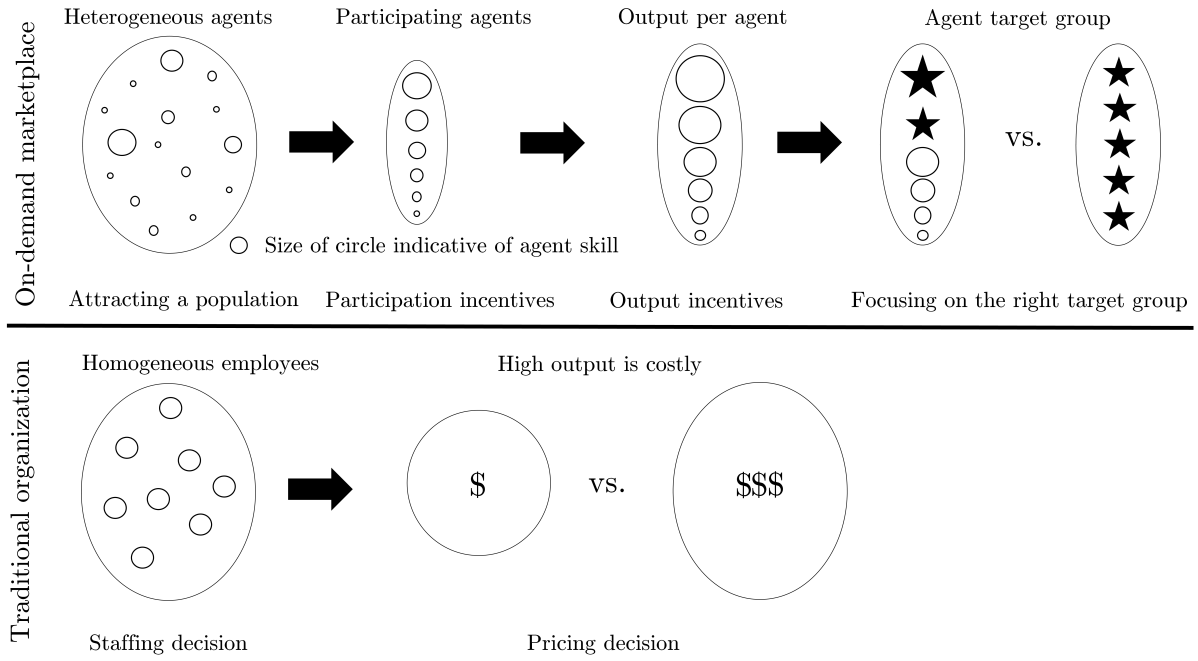


Figure 3.8 An on-demand marketplace vs. a traditional organization.

form of dynamic pricing depending on the output generated by an agent. Interestingly, I show that an over-discriminating pricing scheme can often be detrimental for the efficiency of an on-demand workforce. The optimal mechanism would never discriminate at a key subset of the on-demand agent population, despite their performance discrepancy that may be known to the marketplace! Put differently, *hiding* information is beneficial.

My dissertation research is consisted of three essays and a case study. The first essay studies the participation and effort incentives design of an innovation contest. The second essay concerns the optimal priority classes scheme to maximize profits and welfare of a work-from-home contact center. The third essay extends the previous essays in a dynamic participation model of a product support forum in which outside users and firm staff members compete for service. A brief account of the individual chapters is provided below.

A brief review

Chapter 1 presents a model of an innovation contest. Most papers in the extant contest literature begin with specifying the number of contestants that are assumed identical to the agent population. However, an innovation contest is conducted over the internet which allows firms to tap into the expertise of a heterogeneous, large-scale population of potential contestants (solver population). Depending on their individual preferences, the solvers strategically determine whether to participate into the innovation contest incurring a participation cost (participating solvers).

A first conceptual contribution of Chapter 1 and a key departure from the existing literature is the distinction between participating solvers and potential solvers. Technically, I treat the *ex ante* number of participating solvers as a non-negative random variable who is realized *ex post*

the participation decision of the solvers. Focusing on pure symmetric strategies I show that a Binomial distribution characterizes the number of participating solvers. The participation probability is a decision variable of the solvers who participate by forming symmetric beliefs about the actions of the others. In a self-confirming equilibrium, I require solver beliefs to match the observed outcomes and I show the uniqueness of a pure symmetric equilibrium.

Participating solvers exert effort, which is not observable by the firm. Instead, the firm observes solver *performance* which I model as a convex combination of solver (privately known) ability and (unobservable) effort. The degree to which solver performance depends on solver ability is a structural characteristic of each contest I refer to as "contest specialization". Contest outputs that place a higher weight on solver intrinsic ability (as opposed to his costly effort choice) may encourage solvers to shirk. To fully characterize the effect of contest characteristics on solver participation, I study the impact of contest characteristics on the expected number of participating solvers ("quantitative" measure of participation) and the expected ability of participating solvers ("qualitative" measure of participation).

Next, I study the best way to allocate a fixed firm budget to solvers to optimize the equilibrium *performance* behavior of a subset of the top *participating* solvers (see also Figure 3.8). The highest performer receives the top reward of a pre-announced budget allocation, the second highest performer receives the second highest reward, and so on and so forth. That is, an innovation contest bears a strong resemblance to an auction. In fact, an innovation contest is an *all-pay auction*, since all participating solvers are paying their "bid" (effort) to the "auctioneer" (firm).

The design of an innovation contest presents two unique challenges compared to standard auctions. First, the number and type of participating solvers is only known ex post their entry decision. Further, participating solvers exert effort to compete only with their peers who choose to participate. Second, the effort exerted by each solver is not observed by the firm and is typically influenced by solver ability or luck/feedback that are not revealed to the firm. The techniques proposed in this dissertation chapter, comprising of the use of the self-confirming equilibrium notion and adverse selection arguments, make significant progress in this direction for the optimal innovation contest.

The probabilistic view of the number of participants in a contest constructed in Chapter 1 is applied into Chapter 2 that studies the strategic participation decision of the agents of a work-from-home contact center. Most models in the existing queueing literature either study routing schemes in a fixed number of servers, or the amount of staff members needed is a decision variable of the firm. In contrast, the agents of a work-from-home contact center have a "mind on their own" and they can voluntarily decide whether to work in a pre-specified work shift. That is, ex ante agent participation the number of available agents to route demand to is unknown to the agents and the firm. The firm has to attract agents to work.

In addition, Chapter 2 proposes a focus on priorities on *servers*, as opposed to customer priorities that is the major concern of the queueing literature to date. From a traditional queueing theoretic standpoint, the agents of a work-from-home contact center can be viewed as "virtual customers" who "queue" in order to *provide* service (as opposed to real customers who queue to *receive* service). Given this conceptual connection, I focus on the expected amount of

time that an agent is busy (agent utilization), which can be considered as the "dual" analog of the expected waiting time of a customer. The inherent uncertainty on the number of agents who would choose to participate requires a focus on a relevant performance metric I term the *expected utilization* of an agent.

The voluntary participation choice of the agents in a work-from-home contact center pose three novel research challenges. The first challenge concerns the stability of a service system when the number of its servers is random. As described in the Introduction, I am aware of three different methods to proceed with this issue. I propose a model of endogenous demand for service which guarantees a stable service, while at the same time makes the exact analysis tractable. The second challenge concerns a closed form characterization of the expected utilizations of each agent ranked in an arbitrary priority class. Conditional on the number of participants and using techniques from the traditional $M/M/n$ model, I derive an explicit form for the expected utilization of each priority class and investigate its properties as the parameters of the system change. The third challenge encountered is how to solve for the optimal priority class formation (i.e. *number* of priority classes to form as well as *size* of each priority class). I conceptualize the priority class design of a work-from-home contact center as a contest design problem. Specifically, since agents are paid only on-demand, one can view their expected utilization as their *expected* rewards promised by a contest designer. In addition, I require *all* incoming demand ("budget" to pay the agents) to be satisfied or sent in the queue, which implies a budget constraint to hold for every *realization* of the participating agents (hence, it holds in expectation as well).

My approach to innovation contest design and priority classes partition design of a work-from-home contact center integrates principles from linear programming and game-theoretic mechanism design with probabilistic arguments. Indeed, the optimal innovation and service contest design exhibit some similarities. Specifically, I show that a coarse priority classes partition composed of a few high performers generates the highest profits for the work-from-home contact center. This resembles a generalized form of a winner-take-all scheme recovered in traditional crowdsourcing contests. Employing a limiting argument of the convergence of the Binomial distribution to the Poisson distribution, I show that two priority classes asymptotically maximize social welfare. This is a win-win-win solution for the customers, the agents and the work-from-home contact center.

Finally, Chapter 3 studies the management of a product support forum in which an organization with existing staff members taps into a large online community of "outside" users to serve demand. This is a hybrid, on-demand contact center in which customer service is partially outsourced back to customers. In a product support forum, there is an abundance of service representatives who could potentially provide service on-demand at a negligible cost. However, similarly to an innovation contest or a work-from-home contact center, the focal firm has to provide appropriate incentives for its community users to respond accordingly. Currently, large corporations firms such as Microsoft, Apple, Walmart and PayPal use a virtual reward system of badges to promote fast service of high quality from their communities who engage actively on-demand. This presents to organizations an unprecedented opportunity to provide low cost and high quality service. Given its potential for significant improvements over its traditional

and work-from-home counterparts, I view a product support forum as the *future of the contact center*.

The incentive design of a product support forum resembles a contest for service in which outside users compete with firm servers to resolve customer questions on-demand. However, a product support forum manager does not simply wish to maximize user participation at one question only (as in a single innovation contest). Instead, frequent participation from the community is the objective of a product support forum. Given that easy and hard problems may potentially arrive into the forum, the users and the firm have the option to choose which problems to reply to. Users would only receive the associated reward, if and only if, they arrive before the question-asker abandons service and before the firm. I find that the optimal staffing of a product support forum depends critically on asker impatience. In particular, I show that for sufficiently high impatient askers, the firm should employ a larger workforce whereas for moderately impatient askers, the firm should barely interact into its forum. This provides a game theoretic reason for why Apple prefers to solely rely on its user community while Microsoft devotes some paid capacity to respond, in addition to its respective user community.

Directions for future research

Research on the marketplace design spans across many disciplines and presents a fruitful opportunity for theoretical, empirical and experimental work in the intersection of operations management, compensation, incentives, stochastic systems, and matching theory. I outline below some specific future directions.

Innovation contests with voluntary participation

The theoretical framework regarding solver voluntary participation decision, developed in Chapter 1 can be applied essentially to any variant of the problem already studied in the contest literature so far that assumed an exogenously fixed set of contestants. In particular, innovation contests are conducted in a variety of forms including staged contests, sequential contests and multiple contests that run in parallel. I elaborate further on the last form of innovation contest that is the least studied in the literature to date.

Most of the current literature focuses on the incentive design of a single innovation contest. A model of *multiple*, parallel contests is closer to practice with contest platforms such as Kaggle.com, InnoCentive.com and TopCoder.com. A solver in such a contest platform first decides whether to participate, expending a fixed set-up cost up-front. Conditional on participation, there is a search cost incurred in order to choose *which* contest to participate. Then, a solver exerts effort which depends on his ability type and the actions and types of his competitors. Crucially, a solver only competes with those who chose to participate in the same contest with him, which is only revealed ex post his participation. The anticipated number, ability types, and effort exerted by those solvers would affect the contest choices of the solvers and their effort decisions. Uncertainty about quantity and quality of the solvers may also lead to inefficiencies such as "clustering" of solvers into a few contests that they expect to attract low or moderate competition. Hence, the contest designers need to understand solver behavior in this setting

to design incentives that increase the breadth of search while at the same time increase solver effort.

Further, although many contests run in parallel, they differ in the deadline they set and award structure they offer to attract high performance from outside solvers. To the best of my knowledge, there is only a handful of papers that study the optimal deadline decision of a firm, but none that considers this decision in a competitive setting. Setting a shorter deadline may decrease the quantity of participants but may actually increase the quality of the submitted solutions. On the other hand, setting a long deadline would attract a larger pool of solvers who would compete more fiercely with potential detrimental effects on the output generated by them.

Service systems with random servers

Chapter 2 provides the first queueing model to date in which the number of servers is a random variable and is affected by their strategic participation decisions. I have focused on the special case in which the distribution of participating solvers is Binomial and the arrival and service times are markovian. Future research could study service settings with richer dynamics such as voluntary participation and voluntary exit, time-varying participation rates, and non-markovian service systems with random servers. A deep understanding of the game-theoretical underpinnings that determine the strategic participation decision of such servers is required. Only then, researchers can extend current queueing models into that direction.

Another direction is to establish stochastic fluid approximations for all known service performance metrics of a service system with on-demand capacity. In particular, fluid approximations for the expected waiting time, abandonment probability, queue length and server utilization behave fundamentally different and new theoretical tools should be developed to discover them.

Work-from-home contact centers

A work-from-home contact center is a marketplace of freelancers who compete to serve customers of an organization and can be studied from various angles. In particular, one could study the process of hiring, training, or retaining geographically dispersed agents in the absence of supervisors (flat organizational structure). In a related paper, Yakubovich and Lup (2006) study the practice of recruitment in a work-from-home contact center and show that referrals maintain an important labor market role.

Further, one could compare the service model of a work-from-home contact center with a question-and-answer forum. Specifically, future research could study the organizational boundaries of service systems and whether and when service should be delegated to freelance agents versus members of an online community. All theoretical notions of service quality and service performance metrics can be appropriately defined in such a setting and assumptions made on analytical models can be empirically tested given the abundance of publicly available data of forums such as StackOverflow.

Ride-sharing marketplaces

The incentive design for participation of passengers and drivers of a ride-sharing marketplace is fundamentally different from that of the agents of a work-from-home contact center who enter for the entire duration of a pre-determined work shift. In particular, in ride-sharing the drivers should be motivated to stay long in the platform and participate often. Further, the passengers should be encouraged to use the ride-sharing service frequently. Ways to achieve the latter include the provision of discounts or coupons for use. I argue that it is essential to target a key subset of the customer population that uses the service more often than others. Symmetrically, a marketplace should reward its best independent contractors more than others. Discriminating on the demand *and* the supply side combined has the potential to reap the benefits of both sides.

Competition in ride-sharing marketplaces

Most papers that study two-sided marketplaces to date focus on a single two-sided marketplace, neglecting any competitors who offer alternatives to the drivers and passengers of a marketplace. The nature of competition in this strategic interaction resembles the setting with "multiple innovation contests that run in parallel". Indeed, in both settings the agents can voluntarily choose which "contest" to compete in by making a forecast on the number of other contestants who will also choose similarly. However, the "rewards" in such a contest are stochastic, as each agent does not know ex ante the realized demand of a given marketplace, which is further affected by the service level offered by a competing marketplace.

Multi-sided on-demand marketplaces

Another direction is to extend existing two-sided models to multi-sided marketplaces (MSM) such as LinkedIn. In particular, LinkedIn connects users with potential employers, but brings together other "sides" such as advertisers and HR analysts who target potential workers and LinkedIn advisors. How do these different groups interact? When should an MSM invite another group, and when it should avoid "search" and focus on "breadth" to increase the value generated by its current stakeholders? I leave such directions to future research.

Online product support forums

An immediate extension of the forum model considered in Chapter 3 is an adverse selection model that incorporates abilities of users. Indeed, as noted in Chapter 1 in pure symmetric strategies with homogeneous agents either all agents participate, or no agent participates. However, the possibility of information asymmetry implies that only a subset of the potential users would actually participate. How many and of which type of users should the product support forum manager motivate to participate is then a non-trivial decision.

The self-confirming equilibrium techniques developed in Chapter 1 of this dissertation are directly applicable to characterize the equilibrium behavior of the heterogeneous users of a product support forum viewed as a contact center. Incorporating an expected waiting time term into the objective of the firm would render the exact analysis intractable. That is, a

stochastic fluid approximation would be required to simplify firm's objective. I expect the latter to be challenging since the rate of service provided by the users would depend on the size of user population. Hence, to develop an appropriate stochastic fluid approximation one should first characterize the aforementioned subtle dependence before applying limiting arguments as user population grows without bound and by conditioning on the number of participants. To the best of my knowledge, such stochastic fluid limits in which the participation probability of random servers depends on server population have not been developed yet.

Further, one can even attempt to compare third-party question-and-answer (user-to-user) forums with product support forums in which there is interaction of a *firm*. Whether employing staff members to reply to questions in the presence of an active online community would result in an increase in service and profit is left for future research.

Implications for practitioners and the future of work

On the managerial side, the effective incentive design of marketplaces that leverage outside agents to generate innovation or provide service on-demand requires a deeper understanding of three factors that confound decision making: the amount and intrinsic quality of participating agents, the competitive nature of the marketplace, and the objective of the manager. I provide a theoretical framework that helps understand why some practices and incentive structures who perform well in some marketplaces might require a careful re-design to be applied into a different organizational environment. The interplay between a large number of agents of low skills with a few high performers would demand to boost the incentives at the top in order to screen out low achievers, in order to attract more participants from the top and soften the competition. The right incentive structure would critically depend on whether the marketplace manager is focusing on a short-run objective to increase market share, or on a long-run objective to increase the size of the agent population as well as their capabilities by investing into training. I caution that attracting a large number of participants may significantly deteriorate their expected quality, which may further affect future demand due to a drop in the quality of service offered. The rationale behind these effects is of significant managerial value.

This dissertation focuses on innovative business models for service support and innovation creation that essentially define the future of work. I expect that there is a growing trend for firms to open their boundaries and increasingly adopt crowdsourcing practices in their internal processes. The benefits of tapping into a crowd on-demand come with a number of challenges for key stakeholders that need to be incorporated into the decision-making process. Hence, I expect a similar shift of the academic research from traditional approaches into this emerging research stream. Such a research shift may even shape a new definition of the field of Operations Management towards "matching demand with *crowdsourced* supply".

Appendix

This page is intentionally left blank

Appendix $\mathcal{A}$

LiveOps Inc.: The Contact Center Reinvented[1]

[LiveOps] is changing the idea of what the contact center is. It used to be a cost center, sitting in the back of the building forgotten. You just needed a warm body to answer phones. But that's absolutely not the case now. It's moving to the front office. People are realizing: "Wait a second, they're actually spending more time with actual customers than marketing could ever dream of, so why isn't this part of marketing?" And managers are getting much more: "Give me more data. Give me more reports", so I think the agent needs to become the "super-agent".

Marty Beard, CEO, LiveOps[2]

In September 2005, Harley Jones, Chief Operating Officer of the American Red Cross (ARC), was spending another restless night watching TV. In a decade-long experience of disaster relief, Harley had seen multiple relief operations including Hurricanes Isabel, Charley and Ivan, but none as severe as this one. Hurricane Katrina had recently dissipated, having wrecked the economies of Mississippi and Louisiana, forcing over a million people to leave their homes – the biggest diaspora in the history of the United States[3]. Amid the chaos, the ARC had to help storm evacuees connect with their relatives by setting up a fast and reliable contact center.

First, he needed to make some calls to major service providers and decide on which offer to take. As with most cases handled by the ARC, this was a matter of urgency. The main challenge was how to set up a large enough contact center within hours to be able to respond efficiently in the midst of the disaster. On top of that, the total cost of the operation had to be taken into account as it would be covered by donations and charities.

[1] This Chapter is based on joint work with Karan Girotra and Serguei Netessine; see Stouras et al. (2014).

[2] Jon Xavier, "LiveOps' Marty Beard on why those creepy chat boxes on e-commerce sites are changing customer service forever", Silicon Valley Business Journal, 21 March, 2014.

[3] Anthony E. Ladd, John Marszalek, and Duane A. Gill. The Other Diaspora: New Orleans Student Evacuation Impacts and Responses Surrounding Hurricane Katrina. Retrieved on 10 September 2014.

A.1 The Contact Center Industry

A.1-1 Background

A contact center[4] can be defined as a coordinated system of people, processes, technologies and strategies that provide access to information, resources and expertise through multiple channels of communication such as e-mail, instant messaging chat, live video chat, as well as responding to social media posts, apart from the telephone call option traditionally offered, enabling interactions that create value for the customer and the organization[5]. The majority of large companies use contact centers as a means of managing their customer interaction. They can be operated either by an in-house department responsible for day-to-day communications with customers (inbound), or by outsourcing customer interactions to a third party (outbound).

The global contact center market was worth US$3.4 billion in 2014 and was growing by 3.6% per year. Though growth in mature markets such as North America (with a 22% market share) was fairly flat at just 2.2% per year, Asia Pacific (which accounted for 34% of the market) was growing at a rate of 4.9% annually and accelerating (see Figure A.2 and Figure A.3).

To understand how a contact center functions and the technology involved (see diagram in Figure A.4), consider the case of a customer calling the electricity company about paying a bill. He or she dials a single customer service number. The long-distance or public switched telephone network (PSTN) company carries the call to the contact center's privately-owned switch or private automatic branch exchange (PABX). The call may be connected through the PABX to an interactive voice response (IVR) that queries the customer's needs. If the customer asks to speak to an agent, the call is transferred from the IVR to an automatic call distributor (ACD). The ACD uses the information from the IVR and the customer data server to route the call to a trained agent who can handle bill payments and speak the customer's language. Computer telephony integration (CTI) technology may also be employed to facilitate agent-customer interaction by automatically opening the customer's file on the computer as the agent picks up the call.

For a contact center in a sales environment, one measure of customer service is the amount of time that a customer is on hold (i.e. waiting), which is closely tied to revenue per customer. The longer a customer is on hold, the less excited s/he is about purchasing the product (and the more likely to hang up). To offer superior service by keeping waiting time short, the traditional contact center manager makes staffing decisions based on forecasts of the anticipated call volume (according to time of day/year, weather, industry served, etc). Various metrics are used to evaluate the performance of a contact center, or more generally a service system. One way to assess systems' congestion (the number of jobs fed to agents) is to determine their utilization level – the percentage time spent serving incoming requests at a chosen service rate. In practice, the utilization level is the number of jobs assigned to an agent vs. the number of jobs that the agent can actually do. This varies from 0% (non-busy) to 100% (fully blocked).

Another dimension of the performance of a service provider is the time required for service

[4]Distinct from call centers, that only handle telephone correspondence, contact centers use a combination of media such as telephone, fax, letter, e-mail and increasingly, online live chat to provide an all-encompassing solution to client, and customer contact (see Figure A.1).

[5]Brad Cleveland, "Call Center Management on Fast Forward", Third Edition, ICMI Press, 2012.

delivery. Important metrics include the average time waiting in the queue before receiving service, the service rate of the servers, and the average queue length. Obviously, the higher the service rate of the agents, the lower the time required to handle service requests, implying faster service and shorter queues on average.

A.1-2 Industry evolution and challenges

A major shift in the contact center industry occurred with the advent of internet telephony (VoIP), and the subsequent fall in international calling rates. Entrepreneurial ventures started setting up contact centers in cheaper locations such as the Philippines, India and Eastern Europe, where low overheads and variable costs offered the benefit of reduced costs. In just under a decade (1995-2005), a large number of US-based contact centers were offshored.

Traditional brick-and-mortar (B&M) contact centers required substantial capital investment for bootstrapping and functioning. Besides the initial investment required to build a fail-safe infrastructure (structural as well as telecommunications), they had high operating costs – agent salaries and human resource overheads accounting for a majority of the expense. The bigger challenge, however, was their ability to scale up and down. Business cycles were seldom predictable and organizations needed to ramp contact center activity up or down correspondingly. Inability to scale up activities meant lost opportunities and unhappy customers, whereas a sluggish scale-down implied substantial losses on unwanted overhead. Other concerns included the time required to train contact center associates and the cost of hiring pricey managers to ensure that the quality of service was not compromised.

Using off-shore contact centers (business process outsourcing, or BPO) in many instances provided an unreliable quality of service as they were culturally and geographically distant from the customers served, even if economically viable. US organizations that had started offshoring their business processes often experienced diminishing quality over time. Although the benefits of quick ramp-up and down were touted, they rarely transpired. Customer complaints proliferated – accents were different and cultural nuances further obfuscated the quality of conversations. A survey[6] on the offshoring and outsourcing activities of 150 North American companies and business units from 1998 to 2006 estimated that companies which outsourced customer service saw a drop of 1% to 5% in market capitalization, as measured by the American Consumer Satisfaction Index, as well as a significant decline in service quality and customer satisfaction, depending on the industry they were in.

B&M and BPO contact centers exhibited certain similarities but had distinctive cost structures (see Figure A.5). Both employed workers of a similar age profile. The typical contact operator population was comprised of college students working to support their studies with additional income (often their first job). For most it was regarded as a temporary occupation, resulting in high employee turnover – typically 100% per year.

Training tended to take place in-house, provided by managers who ensured that agents were familiar with the equipment. At the request of the client company outsourcing its operations to a contact center, agents got extra training on site, learning about the culture in order to handle

[6] Jonathan Whitaker and Claes Fornell, "How offshore outsourcing affects customer satisfaction", MIT Sloan Business Review, 2008.

incoming calls.

Although offshore contact centers mitigated staffing issues for the external service provider, and (given their dispersal across various geographic locations) could operate around the clock, they were associated with reduced control over business functions, reduced monitoring of quality assurance, and a need to put policies in place to ensure customer satisfaction.

Agents in offshore contact centers were often unfamiliar with the corporate culture, practices and values, and thus less dedicated to the company, the customers, and to providing a level of service in line with company standards. There was also a concern that confidential or sensitive information was less secure with overseas contact center agents than local agents who had undergone strict background checks.

Over the years, the concept of the contact center had undergone a series of transformations, signaling a shift in the consumer's outlook towards customer care. With advances in technology, there was instant access to information, and the dependency on contact centers decreased gradually. As LiveOps' Chairman and CEO Marty Beard once put it, "Even the term 'contact center is antiquated and inadequate, painting an image of dark, backroom operations and isolated information providers."

Contemporary contact centers are multi-channel "command centers" with the potential to be vibrant communications hubs, offering a more entrepreneurial approach to success. No longer disconnected from other business functions, large-scale contact centers utilized cutting-edge technology and a vast variety of tools such as CRM, predictive routing, interactive voice response (IVR), self-service knowledge-based communities, interaction recording and big data analytics to deliver an exceptional customer experience across multiple communications channels.

Meanwhile, consumers became increasingly 'connected', channel-agnostic, and expected brands to deliver on their promises regardless of the source – be it online, in-store, or via telephone interaction, mobile or social media. A recent study found that while voice communication was used to be the preferred means of interaction with a firm 90% of the time, this had fallen dramatically to 55%, outpaced by new channels like online chat and social media. Brands that delivered a frictionless customer experience across all channels were rewarded with customer loyalty and higher earnings. Indeed, the volume of customer interaction in an absolute sense had stayed steady, so essentially other channels had grown on top of voice, email and SMS. People were interacting to a greater extent with firms through their social media pages, sharing their experiences about a product or brand via customized social networks – in other words, the remaining 45% was something other than voice.

An additional evolution was that consumers frequently used multiple channels to make a complaint. A disappointed purchaser of a malfunctioning product would not simply call the firm in question, but would also email and express his/her complaints on Twitter and Facebook. Traditional contact centers which focused on one channel were thus unable to handle such cases effectively. An integrated approach was needed so that consumers interacted with one entity across all these channels of communication.

A.2 LiveOps, the Home-shore Contact Center

The increasing availability of the internet in the 1990s had profoundly changed the way people live and work. Consumers were able to pay bills online and shop 24/7. They no longer waited for the morning paper since online news was available in real time, and they used online check-in for flights instead of waiting in line. As long as a device with an internet connection was provided, they could talk, share documents and collaborate with anyone, anywhere, anytime. As the Internet became the standard model for connecting geographically dispersed people at the speed of light, a new paradigm for services and eventually work emerged. In this new 'home-shore' model, employees no longer had to be in the office to communicate. Global communities appeared where people shared their expertise to resolve an issue faced by another user, and consumers wrote instant reviews of peer-rated products.

In an era of always-on availability, Steve Doumar and Doug Feirstein saw a way to leverage the power of the internet to create a new business model in one of the world's largest labour-intensive markets: the contact center. In 2000, they founded LiveOps in Fort Lauderdale, Florida, as a solely home agent contact center. Using traditional contact center technology to route calls to agents working from home and connected to the internet, LiveOps solved previously unknown problems such as how to limit background noise, schedule remote workers and maintain agent quality when the management team had no physical contact with them.

LiveOps' business model combined two innovations to meet emerging consumer-driven needs. Firstly, it was a virtual contact center, where 20,000 "live operators" worked as independent contractors from their own homes and on their own schedules. Using cloud-based solutions, new agents could choose when to work, form team meetings, get training and learn from the experience of existing home agents who shared their experience in a private forum. Secondly, it was based on meritocracy, i.e., best-performing agents received the most calls and, as a result, higher earnings. Home-based agents were constantly evaluated through a variety of metrics, including customer feedback, average call-handling time, call resolution rate, overall availability (i.e. how frequently s/he actually chose to work), past professional experience, and their scores on standardized tests set by LiveOps, among others. Thirdly, it was following an integrated approach across the many different channels a customer may have chosen to interact with a client firm of LiveOps. An agent could now easily follow and respond to a customer request using voice, SMS, live chat, email and social media posts on Tweeter and Facebook.

In order to provide consistent service quality, LiveOps pursued a strategy of community management. Instead of command-and-control, the paradigm shifted to social management based on results. Agents who had control of how they worked were inherently more inspired brand ambassadors for other agents, and this carried through to customer satisfaction and outcomes for their clients. As one independent contractor put it in an online forum:

> *"I have to admit I am not at all competitive - but the opportunity to be on this program and view my metrics compared to others is very inviting."*

To qualify as a LiveOps agent, candidates had to pass a series of tests aimed at simulating their reactions when confronting disappointed customers, replying to angry emails or posts made on social media, or patiently helping customers place an order. Having relevant past experience

was desirable, but LiveOps' platform allowed the time necessary for agents to acquire new knowledge and expertise.

In addition, applicants needed a broadband internet connection and a separate phone line at home earmarked for work only. The virtual contact center did not cover the costs of broadband service or separate phone line, but as independent contractors of LiveOps agents could deduct the costs from their taxes.

Over 300 companies and brands around the world, including Salesforce.com, Symantec, Coca-Cola, eBay, Royal Mail Group, Ideal Living, Pizza Hut, Amway and BeachBody.com, relied on LiveOps' technology to ensure effective omni-channel interactions with their customers.

LiveOps was able to handle volume spikes without sacrificing customer service, in one case ramping from zero to 50,000 calls in eight weeks while maintaining average order value sales conversion rates. For another product launch, the virtual service provider handled call volume spikes of 625% while keeping service levels constant. LiveOps' flexible staffing structure compared favourably with the hundreds of agents required for a traditional contact center to replicate such an extraordinary service level.

A.3 Virtual vs. Traditional Contact Centers

Home-shoring a contact center had several advantages over its brick-and-mortar and offshore counterparts (see Figure A.5). Firstly, it attracted a more mature population, such as unemployed parents who were typically well-educated and able to work from home while taking care of children. The fact that agents could choose their work schedule to meet their individual needs made for greater employee satisfaction and much lower turnover rates than traditional contact centers. A survey on LiveOps' distributed workforce reported a 25% to 50% increase in agent job satisfaction and productivity, while turnover could be as low as 4% (compared to the 100% seen at traditional contact center each year)[7].

Secondly, since agents worked remotely, the contact center could provide e-training opportunities to those interested in gaining further knowledge and skills, or in learning from the experience of the other agents shared via a global virtual forum. Such a community allowed knowledge to be easily transferred between employees, minimizing the need for supervisors and managers.

Thirdly, since remote agents were essentially freelancers working from home, a virtual contact center could record all customer interactions, and, using a variety of performance metrics, route calls to the best performers. In contrast, traditional contact centers might have legal restrictions on which conversations they could record – the lines used required a minimum of security as they operated in-house, compared to agents using remote desktops to handle sensitive customer data.

A cloud-based contact center had fewer operating and fixed costs compared to a brick-and-mortar or offshore center. Being able to attract talent across the country, LiveOps could choose from a huge network of 20,000 independent agents, working from wherever they wanted, whenever they wanted, while the company paid them only for the time they were actively engaged

[7] www.liveops.com/engage, accessed on 10 September 2014.

with customers.

In essence, a virtual contact center is a "contact center upside down"[8] (see Figure A.6). Traditional contact centers hired office-based workers who were paid a fixed wage per hour, irrespective of how much demand was routed to the center. This involved operational inefficiencies, as customers faced non-negligible waiting times during peak hours (so higher staffing levels were needed), whiles operators were sitting idle whenever there was a significant drop in demand.

A traditional contact center can be thought of as a build-to-forecast (BTF) production system[9], whereas a virtual contact center is a flexible service provider that is responsive to demand levels in a similar way to a build-to-order (BTO) production line[10]. If demand for the service is high (as in the morning) remote agents will be tempted to work from home in the expectation of proportionally high earnings. The situation is reversed in off-peak hours, when significantly fewer will be keen to work. In this way a virtual contact center manages the risk of not having customers for its employees – and they in turn are willing to tolerate this in exchange for the flexibility and freedom to choose their work schedule and become "CEO of their own destiny" as Maynard Webb, a former CEO of LiveOps, put it[11].

A.4 Challenges for Virtual Contact Centers

While it seemed that LiveOps was 'here to stay', its model of work had some significant challenges that threatened the fundamentals of its organizational existence. Some felt it was walking a thin line with its agent recruiting process. A civil complaint filed on 3 December 2007[12] alleged gross underpayment of salary and benefits for a number of home agents who had worked for LiveOps for over a year, as reported in the Houston Chronicle:

> *"Two agents in Georgia contend they don't even earn the minimum wage when their training time and non-paid downtime between calls are factored in. The two women argue they are employees, not independent contractors, and are entitled to minimum wage and overtime pay."*

Eventually, LiveOps won the lawsuit[13], defending its model of "freedom in the work-space". However, the complaint was filed in Georgia – a state which favored employers –but some experts believed that if such complaints were filed in California or any other liberal state, the outcome could be more severe. Moreover LiveOps was a growing company and thus generally 'off the radar', but how many more such lawsuits might be brought should the company plan to go down the IPO route?

[8] Karan Girotra and Serguei Netessine, "The Risk-Driven Business Model: Four Questions That Will Define Your Company", Harvard Business Review Press, 2014.

[9] Gérard Cachon and Christian Terwiesch, "Matching Supply with Demand: An Introduction to Operations Management", Third Edition, McGraw-Hill/ Irwin, 2012.

[10] *Ibid.*

[11] As cited in an interview of Maynard Webb by Barry Kibrick of Between the Lines available at www.youtube.com/watch?v=RCk9l47ajPE, accessed on 20 September 2014.

[12] http://goo.gl/JJkQ8h and http://goo.gl/p91n5h, retrieved on 11 August 2014.

[13] http://goo.gl/RGJUfM, retrieved on 10 August 2014.

The elimination of fixed operating costs such as infrastructure, buildings and fixed employee benefits and wages created another major challenge for a home-shore service provider: how to attract a sufficient number of agents at a specific time of the day to ensure a high service level for clients. Furthermore, the quality of those choosing to work was also uncertain; a virtual contact center could hardly force agents of its overall workforce population with a specific historical performance evaluation level and experience to become available and work. Logically, if the rules of the system (payment scheme) were properly set, the supply of home-based agents should "follow" the customer demand for the service. By virtue of being able to flexibly adjust the supply of workers to meet demand, this matching is what differentiated LiveOps from its competitors.

A.5 Weighing the options

A brick-and-mortar, offshore contact center, or virtual service provider – which system would be best suited apply to the Katrina crisis? CEO Harley Jones started listing his priorities.

Firstly, efficiently responding would require a significantly large number of agents to allow the ARC offer superior service and negligible waiting time to storm evacuees. They would have to work around the clock, so that an always-on service was provided. There would be no room for dissatisfied customers who could further overload the whole system by calling back to ask for further information – each request had to be resolved by the first call made.

Secondly, the time required to set up such a large-scale contact center was a key component of the relief effort. Jones estimated that an efficient response would require a hotline to be provided in a matter of hours. If advertised through mass media, this could create an unprecedented spike in call volume. He was projecting that several hundred agents would be required to connect storm evacuees with their relatives for over a week, working 24/7, but he was still not sure whether demand would significantly exceed this estimate, rendering the whole contact center of the ARC "uncontactable".

Thirdly, the cost of the whole operation had to be taken into account. The offers of the various service providers under consideration would need approval by the ARC budget branch. Although the disaster had attracted a lot of large last-minute donations and government funds, the ARC still had to allocate funding efficiently, without neglecting its charitable and non-profit organizational nature that had to provide emergency assistance in the event of other disasters as well.

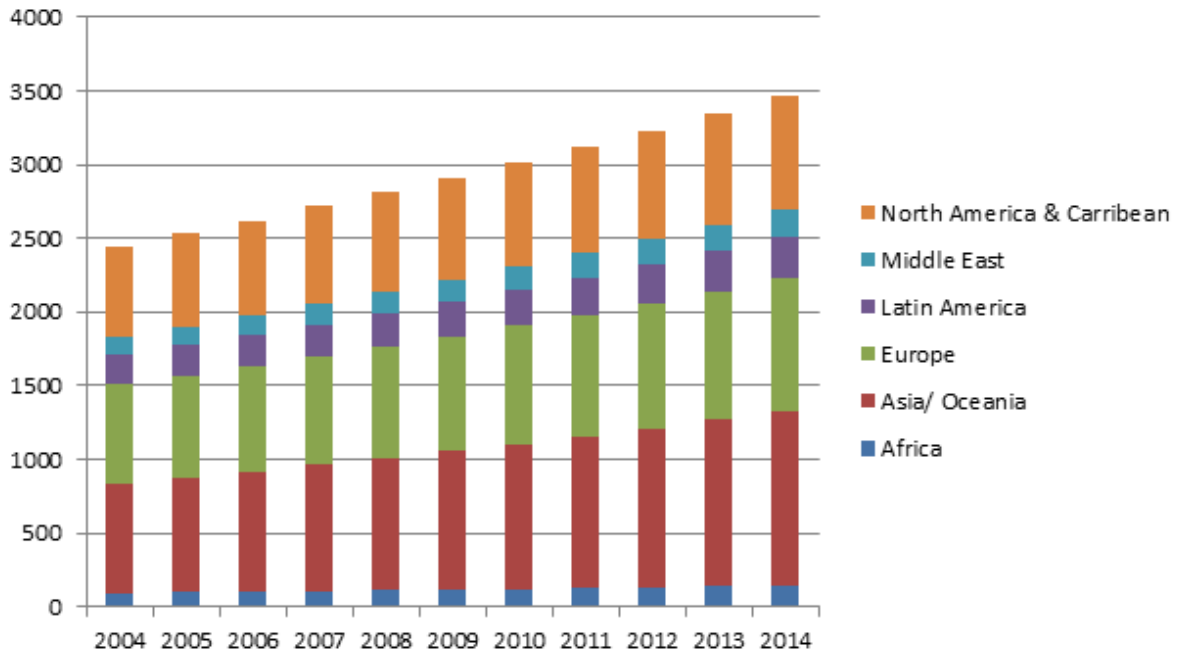
Last but not least, the task to be implemented was fairly simple and did not require any prior technical or professional expertise by the operators. Agents would have to be responsive and to be able to search information gathered by the ARC on 'found' individuals, connecting them with their relatives. The Katrina crisis was unique in the history of the United States. Total property damage and associated costs were estimated to exceed $108 billion (2005), roughly four times the damage wrought by Hurricane Andrew in 1992[14]. Selected agents had to fully understand the gravity of the situation, and thus operators would have to show great sensitivity towards lost family members calling in panic to connect with their close relatives.

[14]http://en.wikipedia.org/wiki/Hurricane_Katrina, accessed on 29 September 2014.

Having dialed the number of a major service provider, Harley Jones was put on hold after an automatic voice reassured him that he had called the right number, followed by a burst of a song by R.E.M. *"The End of the World as We Know It"*. As he held the line, he was struck by how the story of LiveOps had revolutionized the whole service industry.

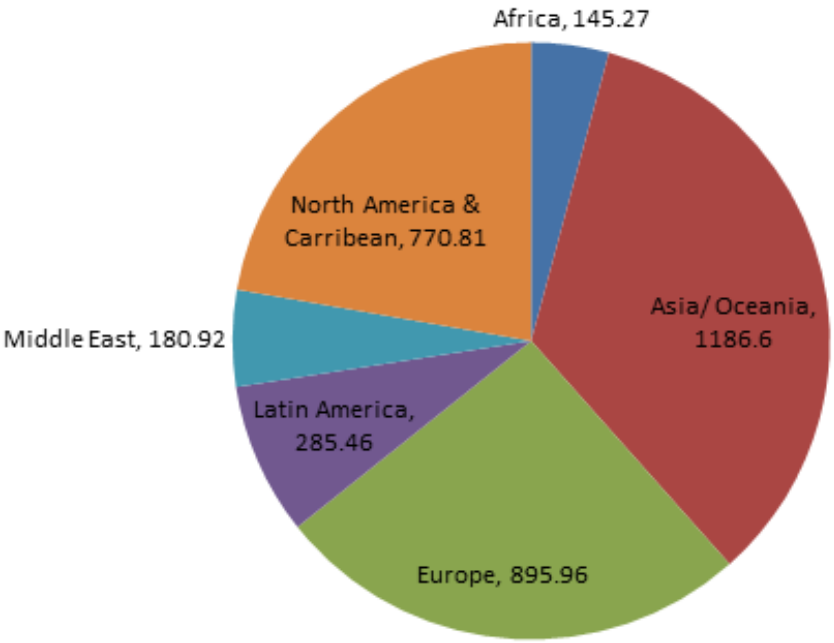
Source: Flickr (http://goo.gl/87LcAg), accessed on 1 September 2014.

Figure A.1 View Inside a Traditional Contact Center.



Source: Consolidated from Parker, P. M. 2010, 'The 2009-2014 Outlook for Contact Centers in Africa', 'The 2009-2014 Outlook for Contact Centers in Asia & Oceania', 'The 2009-2014 Outlook for Contact Centers in Europe', 'The 2009-2014 Outlook for Contact Centers in Latin America', 'The 2009-2014 Outlook for Contact Center in North America & the Caribbean', 'The 2009-2014 Outlook for Contact Centers in The Middle East'.

Figure A.2 Global Market Potential for Contact Centers (US$ Million).

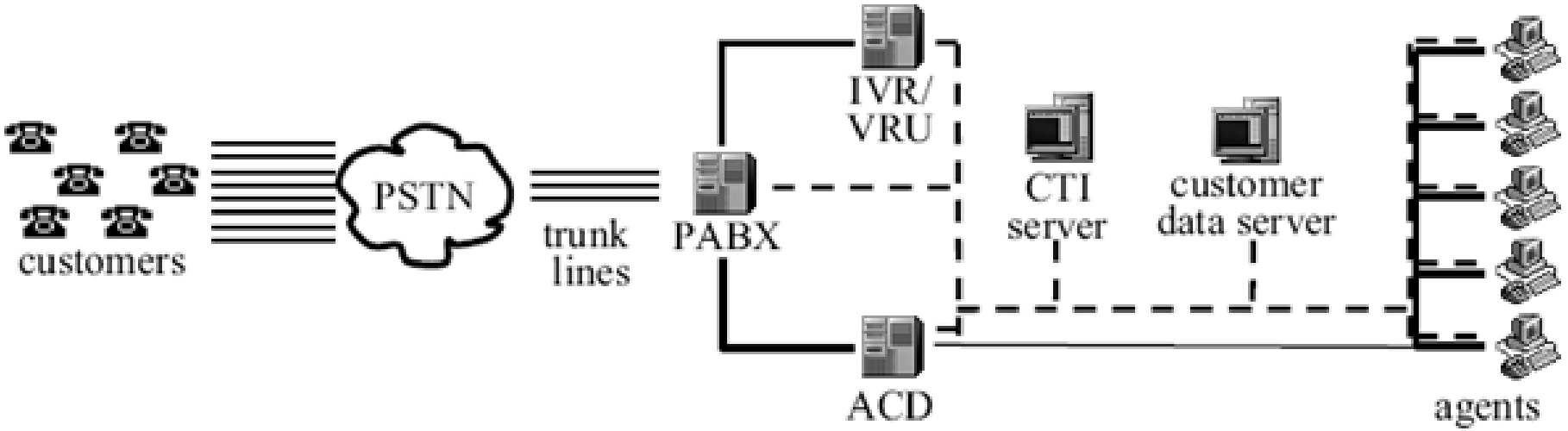


Source: Consolidated from Parker, P. M. 2010, 'The 2009-2014 Outlook for Contact Centers in Africa', 'The 2009-2014 Outlook for Contact Centers in Asia & Oceania', 'The 2009-2014 Outlook for Contact Centers in Europe', 'The 2009-2014 Outlook for Contact Centers in Latin America', 'The 2009-2014 Outlook for Contact Center in North America & the Caribbean', 'The 2009-2014 Outlook for Contact Centers in The Middle East'.

Figure A.3 Global Breakdown of Market Potential for Contact Centers (US$ Million): 2014.

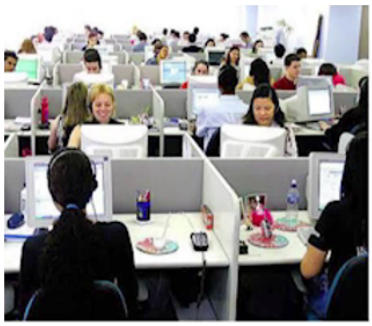
	Agent Profile	Agent training	Security/ Risk	Costs and cons
Brick-and-mortar	Average age 21-24; First job; No college degree; High employee turnover; build-to-forecast (BTF) staffing	Basic; In-room instruction	Minimal security; Monitor agents is easy	Buildings/ Offices; Contact center software and infrastructure; Wages/ Employee benefits;
Offshore	Average age 21-24; First job; No college degree; High employee turnover; BTF staffing	Hiring and training of agents is done by the external service provider	Minimal security; Not bound by laws and regulation of the country operating; Difficult to monitor and enforce	Linguistic/ Cultural differences and decreased customer satisfaction; Time-zone differences; Wages paid to employees; Decreased control over business functions;
Home-shore	Average age 38-41; 15+ yeas of work and life experience; 80% college educated; High employee satisfaction and low labour turnover; build-to-order (BTO) staffing	Comprehensive and training on agents' demand; Strong community support by the other agents	Secure VoIP connection has to be established; 100% call recording; Payment Card Industry (PCI) compliant	Wages to paid to home agents; No geographical restrictions; Quality and experience of agents entering the system may vary;

Figure A.5 A Comparison of Different Types of Contact Centers.



Source: Gans, N., Koole, G., Mandelbaum, A. 2003. Telephone call centers: Tutorial, review and research prospects. Manufacturing and Service Operations Management 5(2) 79-141.

Figure A.4 Schematic Diagram of Call-Center Technology.



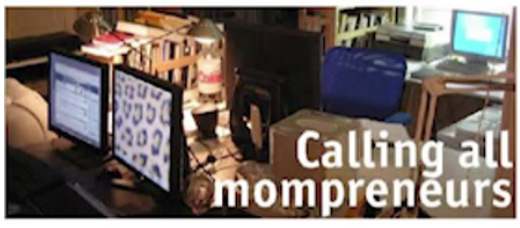
Inshore contact centers

Office-based workers are paid a flat wage and the company bears the risk of call demand volume.

Massive offshore contact centers

Calls are outsourced to service providers in developing countries who bear the demand risk.

liveops®



Home-shore contact centers

Demand risk is mitigated to the home-based employees since they are paid per minute spent on the phone serving a customer, as opposed to the overall time they actually wait for calls to arrive in the system.

Figure A.6 The Evolution of the Risk Profile of the Contact Center.

This page is intentionally left blank

Appendix $\mathcal{B}$

Order statistics

Suppose that $\mathcal{A}_1, \ldots, \mathcal{A}_N$ are N independent and identically distributed (IID) random variables from a common cumulative distribution function (CDF) $F(\cdot)$ and probability density function (PDF) $f(\cdot)$. Following the notation of Shaked and Shanthikumar (2007) let $\mathcal{A}_{j:N}$ denote the random variable corresponding to the jth smallest observation in the random sample $(a_i)_{i=1}^N$, also known as the jth *order statistic*. Further, let $F_{j:N}(\cdot)$ and $f_{j:N}(\cdot)$ denote the CDF and the PDF of the jth order statistic $\mathcal{A}_{j:N}$, respectively. Then,

$$F_{j:N}(a) = \sum_{i=j}^{N} \left(\begin{array}{c} N \\ i \end{array} \right) F^i(a) (1 - F(a))^{N-i} = N \left(\begin{array}{c} N-1 \\ j-1 \end{array} \right) \int_0^{F(a)} t^{j-1} (1-t)^{N-j} \, dt \tag{B.1}$$

$$f_{j:N}(a) = N \left(\begin{array}{c} N-1 \\ j-1 \end{array} \right) F^{j-1}(a) (1 - F(a))^{N-j} f(a) = N \left\{ F_{j-1:N-1}(a) - F_{j:N-1}(a) \right\} f(a), \tag{B.2}$$

for $j = 1, \ldots, N$ and for a in the support of $\mathcal{A}_{j:N}$. A solver is ranked jth highest among N, if and only if, he is ranked $(N-j+1)$th lowest among N. Using differences of order statistics we have

$$\begin{aligned} \mathbb{P}\left[\mathcal{A}_i \text{ ranked } j\text{th highest among } N\right] &= F_{N-j:N-1}(a_i) - F_{N-j+1:N-1}(a_i) \\ &= \left(\begin{array}{c} N-1 \\ j-1 \end{array} \right) F(a_i)^{N-j} (1 - F(a_i))^{j-1} \\ &= \left(\begin{array}{c} N-1 \\ N-j \end{array} \right) F(a_i)^{j-1} (1 - F(a_i))^{N-j}, \end{aligned} \tag{B.3}$$

where we define $F_{0:N}(a_i) := 1$ and $F_{N:N-1}(a_i) := 0$. The third equality follows by the pigeonhole principle.

Next, we state a stochastic order relation among two random variables $\mathcal{X}$ and $\mathcal{Y}$, as well as some useful facts involving order statistics and first order stochastic dominance.

Definition 1 (Usual Stochastic Order)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be two random variables such that

$$\mathbb{P}[\mathcal{X} > x] \leq \mathbb{P}[\mathcal{Y} > x], \quad \text{for all } x \in \mathbb{R}$$

Then $\mathcal{X}$ is said to be *smaller than* $\mathcal{Y}$ *in the usual stochastic order* (denoted by $\mathcal{X} \leq_{st} \mathcal{Y}$).

Equivalently, we have that $\mathcal{X} \leq_{st} \mathcal{Y}$ iff $\mathbb{P}[\mathcal{X} \leq x] \geq \mathbb{P}[\mathcal{Y} \leq x]$, for all $x \in \mathbb{R}$.

Lemma 7 (Shaked and Shanthikumar (2007))**.** *The following relations hold for all x, y in the support of the respective random variable:*

(a) If $\mathcal{X} \leq_{st} \mathcal{Y}$, then $F_{\mathcal{X}}(x) \geq F_{\mathcal{Y}}(y)$ and $\mathbb{E}[\mathcal{X}] \leq \mathbb{E}[\mathcal{Y}]$

(b) $\mathcal{X}_{i:N} \leq_{st} \mathcal{X}_{i+1:N}$, for each $i \in \{1, \ldots, N-1\}$ (or $F_{i:N}(x) \geq F_{i+1:N}(x)$)

(c) $\mathcal{X}_{i:N} \geq_{st} \mathcal{X}_{i:N+1}$, for each $i \in \{1, \ldots, N-1\}$ (or $F_{i:N}(x) \leq F_{i:N+1}(x)$)

(d) $\mathcal{X}_{i+1:N+1} \geq_{st} \mathcal{X}_{i:N}$, for each $i \in \{1, \ldots, N-1\}$ (or $F_{i+1:N+1}(x) \leq F_{i:N}(x)$)

Appendix C

Contest models and their equivalence

In this Chapter, we compare the innovation contest model of "expertise-based projects" of Terwiesch and Xu (2008) with a special case of the model of Moldovanu and Sela (2001) in which cost of effort is convex.

A risk neutral firm announces a way to split her total budget available among any participants out of N agents by offering at most $N-1$ positive rewards R_i such that $\sum_{i=1}^{N-1} R_i$ and $R_i \geq R_{i+1}$ for all i. For simplicity we shall focus on the winner-takes-all case ($R_1 = R$, and $R_i = 0$ for all i), but our argument carries over for any number of rewards.

The agents are endowed with ability a_i which is privately known and drawn from a common knowledge distribution F with $f > 0$ on the support $[a_0, 1]$ with $a_0 > 0$. Each agent may exert effort e_i at a cost $C(e_i)$ and produces *performance* (or output) $x_i = x(a_i, e_i)$. We assume that $C'_e > 0$ and $C(0) = 0$, and $x'_a > 0$ and $x'_e > 0$.

We have the following two contest utility models:

- Expertise-based projects of Terwiesch and Xu (2008). Assume that agent performance is given by $x_i = a_i + \log e_i$ and cost of effort is a linear function $C(e_i) = c \cdot e_i$. We normalize $c = 1$ for simplicity. Then, the utility of agent i is

$$u_{TX}(a_i, e_i) = \begin{cases} R - e_i, & \text{if } a_i + \log e_i = \max_{j \neq i} \{a_j + \log e_j\} \\ -e_i, & \text{else} \end{cases} \qquad \text{(Terwiesch and Xu Model)}$$

Note that the Terwiesch and Xu Model implies that the firm does not observe agent ability neither their effort in order to reward them. Instead, the highest agent ranked by performance receives the award.

- Contest model of Moldovanu and Sela (2001) with convex costs. Assume that agent performance is given by $x_i = e_i$ and cost of effort is a convex function $C(e_i) := \frac{c(e_i)}{a_i}$ with $c'_e > 0$ and $c(0) = 0$. Then, the utility of agent i is

$$u_{MS}(a_i, e_i) = \begin{cases} R - \frac{c(e_i)}{a_i}, & \text{if } e_i = \max_{j \neq i} \{e_j\} \\ -\frac{c(e_i)}{a_i}, & \text{else} \end{cases} \qquad \text{(Moldovanu and Sela Model)}$$

Note that in the Moldovanu and Sela Model agent effort is observable by the firm and the highest agent ranked by effort receives the award.

The next proposition compares these two models and shows that they lead to qualitatively similar agent actions in equilibrium by transforming and re-interpreting the quantities involved.

Proposition 6. *The Terwiesch and Xu Model and the Moldovanu and Sela Model are output equivalent in equilibrium.*

Proof of Proposition 6. We start with the Terwiesch and Xu Model and we note that the ability a_i of agent i is private information to him and is a constant for him. Set $\hat{a}_i := \exp(a_i)$. Then, ranking agents according to $\hat{x}_i = a_i + \log e_i = \log \hat{a}_i + \log e_i = \log(\hat{a}_i\, e_i)$ we have

$$\begin{aligned} u_{TX}(a_i,\, e_i) &= \begin{cases} R - e_i, & \text{if } a_i + \log e_i = \max_{j\neq i}\{a_j + \log e_j\} \\ -e_i, & \text{else} \end{cases} \\ &= \begin{cases} R - e_i, & \text{if } \log(\hat{a}_i\, e_i) = \max_{j\neq i}\{\log(\hat{a}_j\, e_j)\} \\ -e_i, & \text{else} \end{cases} \\ &= \begin{cases} R - \frac{\exp(\hat{x}_i)}{\hat{a}_i}, & \text{if } \hat{x}_i = \max_{j\neq i}\{\hat{x}_j\} \\ -\frac{\exp(\hat{x}_i)}{\hat{a}_i}, & \text{else} \end{cases} \\ &= u_{MS}(\hat{a}_i,\, \hat{x}_i), \end{aligned}$$

where we wrote the effort in Terwiesch and Xu Model as $e_i = \frac{\exp(\hat{x}_i)}{\hat{a}_i}$. The above shows that the Terwiesch and Xu Model can be written in the form of the Moldovanu and Sela Model in which the ability of the agent i is $\hat{a}_i$ and he determines his effort $\hat{x}_i = \hat{x}_i(a_i)$ at a cost of effort $\frac{c(\hat{x}_i)}{\hat{a}_i} = \frac{\exp(\hat{x}_i)}{\hat{a}_i}$. □

Similarly, a multiplicative form for agent output $x_i = a_i \cdot e_i$ can be written as an additive form by taking logarithms and re-interpreting the meaning of "ability", "effort" and "performance". In particular, instead of ranking agents according to $x_i = a_i \cdot e_i$, we rank them according to the additive model $\hat{x}_i = \log x_i = \log(a_i \cdot e_i) = \log a_i + \log e_i = \hat{a}_i + \hat{e}_i$. Relative rankings according to x_i would be preserved by ranking according to $\hat{x}_i$ because the log function is strictly increasing.

A multiplicative form for agent output $x_i = a_i \cdot e_i$ implies that ability and effort may be complements instead of substitutes as in the quasi-linear models of Chapter 1 and Terwiesch and Xu (2008). In a multiplicative model, when an agent exerts zero effort, his output is zero. For example, at Kaggle no contestant can ever win without submitting a (costly) solution to a posted problem, no matter how high ability he may have. As shown above, we can recover this intuition in an additive model in which all participating agents exert strictly positive effort in equilibrium. We refer the reader to the unpublished manuscripts of Erat and Lichtendahl Jr (2015) and Erat and Lichtendahl Jr (2016) for related generalizations of contest models.

Appendix $\mathcal{D}$

Proofs of Chapter 1

D.1 Summary of notation used

Seeker:

- k: (exogenous) number of candidate solutions, or the top solutions that the seeker is interested in ($k \in \{1, \ldots, N\}$).
- R: (exogenous) budget.
- R_i: reward allocated to solver ranked ith out of N and $R_1 \geq R_2 \geq \ldots \geq R_N \geq 0$ (at least one is non-zero).
- w_i: (exogenous) weight of performance ranked ith out of N in seeker's objective for $i = 1, \ldots, N$ and $w_1 \geq w_2 \geq \ldots \geq w_k > 0$.
- m: seeker's choice of awards. It is the number of non-zero awards the seeker splits its available budget into.
- $\Pi_k(m)$: seeker's profit given a choice of awards m for an exogenously fixed k.

Solvers:

- N: (exogenous) number of potential solvers, or size of solver population.
- F: (exogenous) ability distribution with strictly positive density f on its support $[a_0, 1] \subset [0, 1]$.
- $F_{N-m:\,N-1}(\cdot)$ and $f_{N-m:\,N-1}(\cdot)$ the $(N-m)$th lowest out of $N-1$ order statistics CDF and PDF of the ability distribution $F(\cdot)$ respectively.
- a_i: ability of solver i, $i = 1, \ldots, N$. The ability of solver i is a random variable for the seeker denoted by $\mathcal{A}_i$ with distribution F.
- e_i, e_i^*: effort, and equilibrium effort chosen by a solver i respectively.

- u_i: expected utility of solver i.
- $x_i = \gamma\, a_i + (1-\gamma)\, e_i + \varepsilon$: performance of solver i with ability a_i, effort choice e_i and subject to a random shock ε (seeker's subjective taste of a submitted solution).

Other parameters:

- c_p: (exogenous) cost to participate into the contest.
- γ: (exogenous) contest specialization; the degree to which solvers substitute ability for effort, $\gamma \in [0, 1]$.
- $\overline{m}$: upper bound on number of awards m.

D.2 Summary data from innovation contest platforms

In this section, we offer summary statistics for the distribution of innovation contest rewards from data collected[1] from the well known platforms of InnoCentive.com and Kaggle.com, as well as from the popular ideation marketplace of Tongal.com recently analyzed by Kireyev (2016). Most seekers in these platforms were pre-announcing their available budget in advance, as well as their precise way to rank their solvers through a detailed description and were committing to a pre-announced allocation of their budget among the top participating solvers. We note that all these platforms strongly encourage the seekers to guarantee a reward allocation by displaying a prominent message to any prospective seeker, in order to attract high participation and high output from the solvers. Next, we describe our data collection in detail.

First, using a Python script we scrapped all 598 contests that took place on InnoCentive.com during 2012-2015. We find that 192 contests organized offered the entire budget of their respective seeker to the best solution provided by the top participating solver. Table D.1 reports summary statistics for the distribution of rewards and the pre-announced amount of the seekers' budget for InnoCentive.

Second, we use publicly available data of *all* contests that offered a monetary reward on Kaggle.com since its inception on April 2010 until July 2016 (see Table D.2). Such contests include "featured" and "research" contests, as opposed to contests associated with a non-monetary rewards such as "recruitment" contests in which hiring companies are selecting candidates for a data-related position by organizing a contest on Kaggle, or "in class" contests which aim to teach data analysis to novice users.

Lastly, we report a summary of the data collected by Kireyev (2016) from Tongal.com during 2011-2015 (the platform was founded in 2009). We note that, *by design*, all Tongal contests have the *Multiple-Winner* (MW) format we use in §1.2. As Kireyev describes, "all Tongal contests divide rewards evenly among winners. For example, each winning submission receives $250 if a contest offers four rewards with a total budget of $1,000". He also adds that "an important aspect of many contests is that not all participants who consider entering choose to do so". Interestingly, Kireyev empirically shows that on average 78% of the entire solver population participate in a Tongal contest. We reproduce Table 2 of Kireyev (2016) in Table D.3.

[1] All data collected are publicly available to download from the author's website.

Pre-announced contest characteristics	Min	Median	Mean	Max
Number of prizes	1	2	3	15
Total budget of seeker	\$1,000	\$10,000	\$20,987	\$710,000
Number of WTA contests	192 (32%)			
Total number of contests	598			

Table D.1 Data from InnoCentive.com during 2012-2015.

Pre-announced contest characteristics	Min	Median	Mean	Max
Number of prizes	1	3	2	10
Total budget of seeker	\$100	\$10,000	\$26,053	\$500,000
Number of WTA contests	52 (33%)			
Total number of contests	156			

Table D.2 Data from Kaggle.com during 2010-2016 publicly released at goo.gl/NAx4sV.

Pre-announced contest characteristics	Min	Median	Mean	Max
Number of prizes	1	4	5	50
Total budget of seeker	\$500	\$1,000	\$1,450	\$10,000
Participating solvers	58	187	193	499
Solver population per contest	58	235	247	623
Participating solvers/ Solver population per contest	37.23%	78.48%	77.53%	100%

Table D.3 Data from Tongal.com during 2011-2015, as reported in Table 2 of Kireyev (2016).

Overall, the data provided in Table D.1, Table D.2 and Table D.3 show the robust finding that most innovation contests in these platforms offer multiple rewards. In particular, 68% and 67% of the contests organized on InnoCentive and Kaggle offered multiple rewards respectively. In addition, the median number of rewards on InnoCentive was two with a maximum of 15. Similarly, Kaggle seekers pre-announced rewards with a median of three and a maximum of 10. Strikingly, there were seekers who split their available budget with as many as the top 50 participating contestants on Tongal.

D.3 Proofs

Throughout this Chapter, we make extensive use of the following auxiliary result.

Lemma 8 (Maximum Principle)**.** *Let $\mathcal{A}_1$, $\mathcal{A}_2$, ... be a sequence of strictly positive IID random variables and let $\mathcal{N}$ is a discrete random variable with support $\{0, 1, \ldots, N\}$ which is indepen-*

dent of the $\mathcal{A}_i$. Then:

$$\mathcal{A}_{\mathcal{N}:\mathcal{N}} \leq_{st} \mathcal{A}_{N:N}$$

Further, $\mathcal{A}_{\mathcal{N}:\mathcal{N}} = \mathcal{A}_{N:N}$ a.s., if and only if, $\mathbb{P}[\mathcal{N}=0]=0$.

Proof of Lemma 8. Conditional on a arbitrary realization $\{\mathcal{N}=n\}$, due to Lemma 7(d) we have that for all $n \geq 0$: $\mathcal{A}_{n:n} \leq_{st} \mathcal{A}_{N:N}$. Further, condition on a random sample $\{a_1, \ldots, a_N\}$ of (deterministic) size N. We denote the ordered sample as $\{a_{1:N}, \ldots, a_{N:N}\}$, where $a_{N:N}$ denotes the highest observation out of N. Consider a sub-sample of the *top* n observations out of N: $\{a_{N-n+1:N}, \ldots, a_{N:N}\}$. It is immediate that the highest ranked observation out of the *top* n observations is less or equal than the highest ranked observation out of the entire population N. To see that, note that when $n=0$, $\max \oslash = 0$ (by definition) whereas $\max\{a_{1:N}, \ldots, a_{N:N}\} = a_{N:N} > 0$, and that when $n \geq 1$ then $\max\{a_{N-n+1:N}, \ldots, a_{N:N}\} = \max\{a_{1:n}, \ldots, a_{N:N}\} = a_{N:N}$. Then, if $\mathcal{A}_{\mathcal{N}:\mathcal{N}} = \mathcal{A}_{N:N}$ a.s., we have that $\mathbb{P}[\mathcal{N}=0]=0$. Inversely, if $\{\mathcal{N}=0\}$ is a measure zero event, then $\mathcal{A}_{\mathcal{N}:\mathcal{N}} = \mathcal{A}_{N:N}$ a.s. □

The Maximum Principle simply states that the maximum of a top subset of a population of random size is equal to the maximum of the entire population, if and only if, the top subset has size greater or equal to one with probability one.

Proof of Lemma 1. (a) Suppose that the seeker has chosen to split her budget into $m=N$ rewards of value $\frac{R}{N}$ each to anyone who chooses to participate. Then, no solver exerts any effort in equilibrium, since exerting effort is costly and it would not increase his chances of getting a higher expected revenue.

Next, we analyze the participation behavior of the solvers focusing on symmetric equilibria. Assume that solver i participates with probability p_i while all other $N-1$ solvers participate with a best response probability $p = b(p_i)$ that is assumed differentiable. The seeker does not have to spend her entire budget (since $\sum_{j=1}^{N} R_j \leq R$) and would allocate as many equal rewards as the endogenously determined participants. Hence, the utility of solver i depends on how many other solvers $\mathcal{K}$ decide to enter: $\frac{R}{1+\mathcal{K}} - c_p$, where $\mathcal{K} \sim \text{Binomial}(N-1, b(p_i))$ due to symmetry. Hence, the expected utility of solver i is the difference between his expected earnings and his participation cost:

$$u_i(p_i) = p_i \cdot R - p_i \cdot c_p - (N-1)\, b(p_i) \cdot c_p \tag{D.1}$$

For p_i to be a best response, $u_i(p_i)$ must be maximized at p_i. The FOC wrt p_i gives $R - c_p - (N-1)\, b'(p_i) \cdot c_p = 0$. Since we are seeking for a symmetric equilibrium, we substitute $p_i = p$ which implies:

$$p^* = \frac{R - c_p}{(N-1)\, c_p} \tag{D.2}$$

This is the best response for a solver, if and only if, $b(p_i) \in [0,\, 1]$. Therefore, solver participation probability is given by $\min\left\{\max\left\{0,\, \frac{R-c_p}{(N-1)c_p}\right\},\, 1\right\}$. It is immediate that in pure and symmetric strategies either all, or none of the N solvers participate.

(b) Allocating $\overline{m} = \max\left\{n \in \{1, \ldots, N\} :\ \frac{R}{n} \geq c_p\right\}$ equal rewards of size $\frac{R}{\overline{m}}$ induces $\overline{m}$ solvers to participate and exert zero effort in equilibrium as their expected earnings are not affected by their choice of costly effort. Subject to the condition $R < N\, c_p$, no solver participates for $m > \overline{m}$

rewards since he can not cover his participation cost. If $m < \overline{m}$, then an ex ante uncertain number of $\mathcal{N}$ solvers participate with a probability $p < 1$. Hence, $\mathcal{N} = \sum_{i=1}^{N} \mathbb{1}_{\{\text{solver } i \text{ participates}\}}$ is a sum of independent Bernoulli trials with success probability p. By definition, we have that $\mathcal{N} \sim \text{Binomial}\,(N,\, p)$. □

Proof of Theorem 1. Consider a population of N solvers and suppose that the seeker has decided to allocate a reward R_j to an solver whose performance is ranked jth among N. We prove the existence and uniqueness of a symmetric pure BNE for the general case where the rewards satisfy $R_j \geq R_{j+1}$ for $j = 1, \ldots, N-1$ and there are at least two different rewards such that $\sum_{j=1}^{N} R_j \leq R$. We specialize to the sub-class of contests with MW format in the end of the proof.

We introduce some notation. All N solvers simultaneously make participation probability and effort actions, which are denoted by $\boldsymbol{\alpha} = (p_i,\, e_i) \in \{0,\, 1\} \times [0,\, +\infty)$. Let $\boldsymbol{x} = (x_1, \ldots,\, x_i, \ldots,\, x_N)$ and $\boldsymbol{e} = (e_1, \ldots,\, e_i, \ldots,\, e_N)$ denote the performance and effort vectors respectively. Let $\boldsymbol{\phi} = (\phi_1, \ldots,\, \phi_i, \ldots,\, \phi_N)$ denote the participation status of the solvers, where $\phi_i \in \{0,\, 1\}$. Then, solver i participates, if and only if, $\phi_i = 1$, and does not participate otherwise. It is crucial to note that solver i competes *only* with the number of other solvers who choose to participate, and that this number is not known to him. Let Φ_{-i} denote the set of all possible $\boldsymbol{\phi}_{-i}$. Further, conditional that solver i participates, the (realized) number of participating solvers is $n\left(\boldsymbol{\phi}_{-i}\right) = 1 + \sum_{j \neq i} \phi_j$, and are not know to solver i ex ante participation.

Conditional on an action vector $\boldsymbol{\alpha} = (\alpha_1, \ldots,\, \alpha_i, \ldots,\, \alpha_N)$, the expected utility of a participating solver i is

$$u_i\left(\alpha_i;\, \boldsymbol{\alpha}_{-i}\right) = \sum_{\boldsymbol{\phi}_{-i} \in \Phi_{-i}} \left\{ \left(\prod_{j \neq i} p_j^{\phi_j} \left(1 - p_j\right)^{1 - \phi_j} \right) \left(\sum_{j=1}^{n\left(\boldsymbol{\phi}_{-i}\right)} R_j \cdot \mathbb{P}\left[x_i^* \text{ ranked } j\text{th out of } n\left(\boldsymbol{\phi}_{-i}\right) \right] \right) \right\} - \frac{e_i^*}{a_i} - c_p, \tag{D.3}$$

where solver i receives reward R_j, if and only if, he is ranked jth highest among the $n\left(\boldsymbol{\phi}_{-i}\right)$ participating solvers.

Assume that in equilibrium the performance of agent i, x_i^*, is strictly increasing in his ability a_i. We prove that our assumption is true in the proof of Theorem 2. Then, the expected utility of agent i is also strictly increasing in his ability a_i, and is further continuous and defined on the compact set $[a_0,\, 1]$. Hence, all participating solvers receive strictly positive expected utility, expect of a unique solver with ability a_{min} (which is an event of measure zero). We refer to the solver with ability a_{min} as the "marginal participating solver" who expects zero utility and hence is indifferent in participating or not, and exerts zero effort.

Focusing on symmetric pure equilibria, solver i conjectures that all other solvers participate with participation probability $\tilde{p}_{-i}$. Given our assumption that solver expected utility is strictly increasing in ability, each of the rest $N-1$ solvers participate and are ranked higher than the marginal solver with probability $1 - F\left(a_{min}\right)$, for all i. That is, in a symmetric BNE we require all solvers conjecture the same participation probability $\tilde{p}$, which does not depend on solver i's ability and must be equal to the actual participation probability, i.e. we have $\tilde{p} = 1 - F\left(a_{min}\right)$.

Further, our assumption that solver expected utility is strictly increasing in ability implies

that if an agent with ability a_j participates, all other agents with higher abilities participate as well. That is, only a subset of random size of the top ability solvers participate. Then, the expected utility of a participating solver is written as

$$u_i(\alpha_i;\, \boldsymbol{\alpha}_{-i}) = \sum_{n=0}^{N} \left\{ \begin{pmatrix} N \\ n \end{pmatrix} \tilde{p}^n (1-\tilde{p})^{N-n} \left(\sum_{j=1}^{n} R_j \cdot \mathbb{P}\left[x_i^* \text{ ranked } j\text{th out of the top } n\right] \right) \right\} - \frac{e_i^*}{a_i} - c_p,$$

Also, we *assume* (and show subsequently on Proposition 2) that the event that no solver participates $\{\mathcal{N} = 0\}$ is of measure zero, which implies that

$$\mathbb{P}\left[x_i^* \text{ ranked } j\text{th out of the top } n\right] = \mathbb{P}\left[x_i^* \text{ ranked } j\text{th out of the entire population } N\right]$$

for all realizations of the number of participating agents $n \leq N$ and for all $j = 1, \ldots, n$ due to Lemma 8.

Hence, the expected utility of a participating solver is simplified to

$$u_i(\alpha_i;\, \boldsymbol{\alpha}_{-i}) = \sum_{j=1}^{N} R_j \cdot \mathbb{P}\left[x_i^* \text{ ranked } j\text{th out of } N\right] - \frac{e_i^*}{a_i} - c_p \tag{D.4}$$

Next, we characterize the marginal participating solver who has ability a_{min} and exerts effort $e^*(a_{min}) = 0$. The expected utility of the marginal participating solver when each of the rest $N-1$ solvers participate with probability $\tilde{p} = 1 - F(a_{min})$ and rank ahead of him is given by

$$\begin{aligned} u(a_{min},\, 0;\, \tilde{p}) &= \sum_{j=1}^{N} R_j \cdot \left\{ \begin{pmatrix} N-1 \\ j-1 \end{pmatrix} (1-\tilde{p})^{N-j} \tilde{p}^{j-1} \right\} \\ &= \sum_{j=1}^{N} R_j \cdot \left\{ F_{N-j:N-1}(a_{min}) - F_{N-j+1:\,N-1}(a_{min}) \right\} \\ &= \sum_{j=1}^{N} (R_j - R_{j+1}) \cdot F_{N-j:\,N-1}(a_{min}) \\ &= c_p, \end{aligned} \tag{D.5}$$

where we define $R_{N+1} := 0$. Since $F(\cdot)$ is *strictly* increasing on $[a_0, 1]$, the order-statistics distribution $F_{N-j:N-1}(\cdot)$ is also strictly increasing on $[a_0, 1]$ for any rank $j = 1, \ldots, N-1$. Hence, the expected utility $u(\cdot,\, 0;\, \tilde{p})$ is continuous and strictly increasing. Also, observe that the terms in brackets in $u(a_{min},\, 0;\, \tilde{p})$ are the PDF of the Binomial$(N-1,\, \tilde{p})$ distribution. That is, $\sum_{j=1}^{N} R_j \cdot \left\{ \begin{pmatrix} N-1 \\ j-1 \end{pmatrix} (1-\tilde{p})^{N-j} \tilde{p}^{j-1} \right\}$ is the expected reward related to this Binomial distribution with $R_j \geq R_{j+1}$. By a coupling argument $u(a_{min},\, 0;\, \tilde{p})$ is strictly decreasing in $\tilde{p}$. Assuming differentiability this implies that

$$\frac{\partial u}{\partial \tilde{p}}\left(a_{min}(\tilde{p}), \tilde{p}\right) < 0 \tag{D.6}$$

Let a_{min} be the solution (if it exists) to the equation $u(a,\,0;\,\tilde{p}) = c_p$. If $u(a,\,0;\,\tilde{p}) > c_p$ for all $a \in [a_0, 1]$, we set $a_{min} := a_0$; if $u(a,\,0;\,\tilde{p}) < c_p$ for all $a \in [a_0, 1]$, we set $a_{min} := 1$. The strict monotonicity of the expected utility $u(\cdot,\,0;\,\tilde{p})$ implies that if a solver with ability a participates, all other solvers with abilities $a' \geq a$ participate as well. Due to symmetry there is a unique (global) participation threshold $a_{min} = a_{min}(m;\,\tilde{p})$ which is the same across all solvers.

In the following two steps, we show that the solvers have a *unique* Bayes-Nash equilibrium belief defined as $\tilde{p} = 1 - F(a_{min})$.

Step 1: The unique participation threshold $a_{min} = a_{min}(\tilde{p})$ is strictly decreasing in the belief $\tilde{p}$. Indeed, the strict monotonicity of solvers' expected utility implies that $\frac{\partial u}{\partial a_{min}}(a_{min},\,\tilde{p}) > 0$ for all $a \in [a_0,\,1]$; hence, for $a = a_{min}$ we have $\frac{\partial u}{\partial a_{min}}(a_{min},\,\tilde{p}) > 0$. We have also shown in (D.6) that $\frac{\partial u}{\partial \tilde{p}}(a_{min}(\tilde{p}),\,\tilde{p}) < 0$. The Envelope Theorem

$$\frac{du}{d\tilde{p}}(a,\,\tilde{p})\bigg|_{a=a_{min}} = \frac{\partial u}{\partial a_{min}}(a_{min}(\tilde{p}),\,\tilde{p}) \cdot \frac{\partial a_{min}}{\partial \tilde{p}}(\tilde{p}) + \frac{\partial u}{\partial \tilde{p}}(a_{min}(\tilde{p}),\,\tilde{p}) = 0$$

implies that $\frac{\partial a_{min}}{\partial \tilde{p}}(\tilde{p})$ must be positive.

Step 2: There exists a unique solution $\tilde{p}^* \in [0, 1]$ to the equation $\tilde{p} = 1 - F(a_{min}(\tilde{p}))$. Indeed, since F is assumed continuous and strictly increasing in ability, the function $k(\tilde{p}) := 1 - F(a_{min}(\tilde{p})) - \tilde{p}$ is continuous and strictly decreasing in $\tilde{p} \in [0, 1]$. If $\tilde{p} = 0$, then anyone who enters will cover his participation cost with positive probability and hence $a_{min}(0) > 0$ and $k(0) = 1 - F(a_{min}(0)) > 0$. If $\tilde{p} = 1$, then this can not be a BNE since in that case the Budget Condition implies that everyone who enters does not cover his participation cost. So, *ex post* the decision to participate is not rational. Hence, if $\tilde{p} = 1$, $a_{min}(1) < 1$ and $k(1) = -F(a_{min}(1)) < 0$. By the Intermediate Value Theorem, a unique $\tilde{p}^*$ must exist such that $k(\tilde{p}^*) = 0$.

The above characterize solvers' participation strategy in the general, weakly monotone allocation of rewards that satisfy $R_j \leq R_{j+1}$ for all $j = 1, \ldots, N-1$. In the special case of MW contests, we substitute $R_j := \frac{R}{m}$ for $j \in \{1, \ldots, m\}$ and $R_j := 0$ for $j \geq m+1$ into $\sum_{j=1}^{N}(R_j - R_{j+1}) \cdot F_{N-j:\,N-1}(a_{min}) = c_p$ and we find that the marginal solver a_{min} is the unique solution of the equation

$$\frac{R}{m} \cdot F_{N-m:\,N-1}(a_{min}) = c_p \tag{D.7}$$

and solver unique participation probability satisfies $p = 1 - F(a_{min})$. □

Proof of Theorem 2. Similarly to the proof of Theorem 1 we first solve the general case where the rewards satisfy $R_j \leq R_{j+1}$ for $j = 1, \ldots, N-1$ and there are at least two different rewards such that $\sum_{j=1}^{N} R_j \leq R$. If the contest specialization satisfies $\gamma = 1$, then exerting effort does not improve the performance ranking of a solver and all participating solvers exert zero effort in equilibrium: $e^*(a) = 0$ for all $a \geq a_{min}$. Next, we find solvers' equilibrium effort when $\gamma < 1$.

Suppose that a participating solver i chooses performance x_i and that all other participating solvers choose a performance level according to the best response function x^* which is *assumed* to be strictly increasing and differentiable in the ability type a_i of solver i. The expected utility

of solver i is given by D.4.

By the law of total probability the (unconditional) probability that participating solver i wins the top reward R_1 is

$$\begin{aligned}\sum_{k=0}^{N-1} \mathbb{P}\left[x_i \text{ ranked 1st among top } k\right] \cdot \mathbb{P}\left[\mathcal{K}=k\right] &= \sum_{k=0}^{N-1} \mathbb{P}\left[x_i \text{ ranked 1st among } N\right] \cdot \mathbb{P}\left[\mathcal{K}=k\right] \\ &= \mathbb{P}\left[x_i \geq x^*\left(\mathcal{A}_i\right)\right]^{N-1} = \mathbb{P}\left[\mathcal{A}_i \leq \left(x^*\right)^{-1}\left(x_i\right)\right]^{N-1} \\ &= F\left(\left(x^*\right)^{-1}\left(x_i\right)\right)^{N-1}\end{aligned}$$

Trivially, if solver i encounters no other opponents upon entry he exerts zero effort and wins the top reward R_1. Note that the ability of solver i is not known to the rest solvers, denoted by the random variable $\mathcal{A}_i$. Similarly, the probability that solver i receives the second reward R_2 is

$$\begin{aligned}\mathbb{P}\left[x_i \text{ ranked 2nd among } N\right] &= (N-1) \cdot \mathbb{P}\left[x_i < x^*\left(\mathcal{A}\right)\right] \cdot \mathbb{P}\left[x_i \geq x^*\left(\mathcal{A}\right)\right]^{N-2} \\ &= (N-1) \cdot \left(1 - F\left(\left(x^*\right)^{-1}\left(x_i\right)\right)\right) \cdot \left(F\left(\left(x^*\right)^{-1}\left(x_i\right)\right)\right)^{N-2} \\ &= F_{N-2:\,N-1}\left(\left(x^*\right)^{-1}\left(x_i\right)\right) - F_{N-1:\,N-1}\left(\left(x^*\right)^{-1}\left(x_i\right)\right)\end{aligned}$$

By the expression (B.3) in Appendix B we have that

$$\begin{aligned}\mathbb{P}\left[x_i \text{ ranked } j\text{th among } N\right] &= \left(\begin{array}{c} N-1 \\ N-j \end{array}\right) F\left(\left(x^*\right)^{-1}\left(x_i\right)\right)^{j-1} \left(1 - F\left(\left(x^*\right)^{-1}\left(x_i\right)\right)\right)^{N-j} \\ &= F_{N-j:\,N-1}\left(\left(x^*\right)^{-1}\left(x_i\right)\right) - F_{N-j+1:\,N-1}\left(\left(x^*\right)^{-1}\left(x_i\right)\right)\end{aligned} \tag{D.8}$$

Solver i determines his performance level x_i solving

$$\max_{x_i \geq \gamma \cdot a_i} \sum_{j=1}^{N} \left(R_j - R_{j+1}\right) \cdot F_{N-j:\,N-1}\left(\left(x^*\right)^{-1}\left(x_i\right)\right) - \frac{1}{1-\gamma}\,\frac{x_i - \gamma \cdot a_i}{a_i}, \tag{D.9}$$

Observe that if

$$\max_{1 \leq j \leq N} \left(R_j - R_{j+1}\right) < \frac{1}{(1-\gamma) \cdot a_i}$$

or equivalently for a sufficiently high contest specialization

$$\gamma > \hat{\gamma} := 1 - \frac{1}{a_0 \cdot \max_{1 \leq j \leq N}\left(R_j - R_{j+1}\right)} \tag{Gamma Condition}$$

choosing $x_i^* := \gamma \cdot a_i$ leads to positive expected utility (conditional on participation), while choosing any $x > x_i^*$ results in negative expected utility. Hence, when the Gamma Condition is satisfied, solver i chooses $x_i^* = \gamma \cdot a_i$, i.e. he exerts *zero* effort in equilibrium.

Fix a contest specialization $\gamma \in [0,\, \hat{\gamma}]$. The FOC of solvers' expected utility (D.9) wrt x_i gives

$$\sum_{j=1}^{N} \left(R_j - R_{j+1}\right) \cdot f_{N-j:\,N-1}\left(\left(x^*\right)^{-1}\left(x_i\right)\right) \cdot \frac{1}{\left(x^*\right)'\left(\left(x^*\right)^{-1}\left(x_i\right)\right)} = \frac{1}{(1-\gamma)\; a_i}$$

At the symmetric equilibrium, $x_i = x^*(a_i)$ and the above condition becomes

$$\sum_{j=1}^{N} (R_j - R_{j+1}) \cdot f_{N-j:\,N-1}(a_i) \cdot \frac{1}{(x^*)'(a_i)} = \frac{1}{(1-\gamma)\, a_i}$$

$$(x^*)'(a_i) = (1-\gamma) \sum_{j=1}^{N} (R_j - R_{j+1}) \cdot f_{N-j:\,N-1}(a_i) \cdot a_i \tag{D.10}$$

Imposing the boundary condition of the "marginal participating solver" $x^*(a_{min}) = \gamma\, a_{min}$ on (D.10) and integrating both sides from a_{min} to a_i we have

$$x^*(a_i) = \gamma\, a_{min} + (1-\gamma) \sum_{j=1}^{N} (R_j - R_{j+1}) \cdot \int_{a_{min}}^{a_i} a \cdot f_{N-j:\,N-1}(a)\, da \tag{D.11}$$

Since $R_j - R_{j+1} \geq 0$ for at least two rewards, the expression (D.11) verifies that $x^*(a_i)$ is strictly increasing in a_i and continuously differentiable as initially assumed. Note that the strict monotonicity of $x^*(\cdot)$ implies that ties in solver performance ranking (which are broken uniformly at random) are events of measure zero, due to the atomless assumption on the ability distribution F.

Using $x^*(a_i) = \gamma\, a_i + (1-\gamma)\, e^*(a_i)$ we get the equilibrium effort for $\gamma \in [0,\, \hat{\gamma}]$

$$e^*(a_i;\, \gamma,\, m) = \frac{\gamma}{1-\gamma}\,(a_{min} - a_i) + \sum_{j=1}^{N} (R_j - R_{j+1}) \cdot \int_{a_{min}}^{a_i} a \cdot f_{N-j:\,N-1}(a)\, da \tag{D.12}$$

By the expression (D.12) we have that the equilibrium effort $e^*(a_i;\, \gamma,\, m)$ of solver i is non-monotone in his ability a_i. Further, the definition of $\hat{\gamma}$ implies that participating solvers exert positive equilibrium effort, i.e. $e^*(a;\, \gamma) > 0$ for all $a \geq a_{min}(m)$.

Next, we check that the sufficient second-order condition (SOC) is satisfied. The derivative of the expected utility of solver i wrt his effort $u_e(a_i,\, e)$ is non-negative for all $e < e^*(a_i)$ and non-positive for all $e > e^*(a_i)$. Thus, the expected utility $u(a_i,\, \cdot)$ is pseudo-concave and it is maximized at $e^*(a_i)$ given by (D.12).

Finally, we have initially considered the general case where the reward allocation satisfies $R_j \leq R_{j+1}$ for $j = 1, \ldots, N-1$. In a MW allocation, we substitute $R_j = \frac{R}{m}$ for $j \in \{1, \ldots, m\}$ and $R_j = 0$ for $j \geq m+1$. Then, $\max_{1 \leq j \leq N} (R_j - R_{j+1}) = \frac{R}{m}$ and the Gamma Condition is satisfied when $R < \frac{1}{a_0 \cdot (1-\gamma)}$. Set $\hat{\gamma} = \left(1 - \frac{1}{a_0 \cdot R}\right)^+ < 1$. For any arbitrary contest specialization $\gamma \in [0,\, \hat{\gamma}]$ the equilibrium effort in the special case of MW contests is

$$e^*(a_i;\, \gamma,\, m) = \begin{cases} \frac{\gamma}{1-\gamma}\,(a_{min}(m) - a_i) + \frac{R}{m} \cdot \int_{a_{min}(m)}^{a_i} a \cdot f_{N-m:\,N-1}(a)\, da, & a_i \geq a_{min} \\ 0, & a_i < a_{min} \end{cases} \tag{D.13}$$

Further, for any arbitrary contest specialization $\gamma \in (\hat{\gamma},\, 1]$ we have that $e^*(a;\, \gamma) = 0$ for all participating solvers with ability $a \in [a_{min},\, 1]$. □

Proof of Proposition 2. We first show that $R > c_p$ is a necessary and sufficient condition to guarantee that at least one solvers participate. Equivalently, we show that $\mathbb{P}[\mathcal{N} \geq 1] = 1$, if

and only if, $R > c_p$. For sufficiency, assume that $R > c_p$. Denote by

$$P_j(p) = \left(\begin{array}{c} N-1 \\ j-1 \end{array} \right) (1-p)^{N-j} p^{j-1}$$

For any weakly monotone allocation of rewards $(R_j)_{j=1}^N$ let $p_0 \in (0,\, 1)$ be the unique solution to the equation $\sum_{j=1}^N R_j \cdot P_j(p) = c_p$. Such a contest must induce solvers to enter with probability higher than p_0 in a symmetric equilibrium. Suppose towards a contradiction that all solvers enter with probability $p \in (0,\, p_0)$. This cannot be an equilibrium as solver i has a profitable deviation. Indeed, exerting zero effort while all other solvers do not participate yields utility $u_p = \sum_{j=1}^N R_j \cdot P_j(p) - c_p > 0$. Hence, his *expected* utility is $u_p \cdot (1-p)^{N-1} + 0 \cdot \left(1 - (1-p)^{N-1}\right) > 0$, a contradiction to the definition of the symmetric equilibrium. For necessity, assume that there exists a feasible reward allocation $(R_j)_{j=1}^N$ that induces $\mathcal{N} \geq 1$ to participate w.p. 1 when $c_p > R$. Then, $\sum_{j=1}^N R_j \cdot P_j(p) - c_p \leq R - c_p < 0$, a contradiction.

Next, we show that $R < N\, c_p$ is a necessary and sufficient condition to guarantee that at most $N-1$ solvers participate. Equivalently, we show that $\mathbb{P}[\mathcal{N} \leq N-1] = 1$, if and only if, $R < N\, c_p$. For sufficiency, assume that $R < N\, c_p$. By the definition of $\overline{m}$ we have that allocating $\overline{m}$ equal rewards results in $\overline{m} < N$ agents to participate w.p. 1 since $\frac{R}{\overline{m}} - c_p = 0$. For necessity, assume that there exists a feasible reward allocation $(R_j)_{j=1}^N$ that induces $\mathcal{N} \leq N-1$ to participate w.p. 1 when $R \geq N\, c_p$. Then, allocating N equal rewards of value $\frac{R}{N}$ induce all N solvers to enter w.p. 1 and exert zero effort. Hence, $\mathbb{P}[\mathcal{N} \leq N-1] = 0$, a contradiction.

Finally, the condition $R > \frac{1}{a_{min}}$ is a sufficient and necessary condition for solvers to exert strictly positive effort upon entry (as shown in Theorem 2). Hence, imposing $R > \frac{1}{a_0} > \frac{1}{a_{min}}$ is a sufficient condition for this to happen. □

Proof of Corollary 1. (a) Fix an $a_{min} \in [a_0,\, 1]$ and let $\Delta(m) := \frac{1}{m} \cdot F_{N-m:\, N-1}(a_{min}) - \frac{1}{m+1} \cdot F_{N-(m+1):\, N-1}(a_{min})$. Using the integral representation of order statistics (B.1) we have:

$$\begin{aligned} \Delta(m) = & \frac{N}{N-m} \left(\begin{array}{c} N-2 \\ m-1 \end{array} \right) \int_0^{F(a_{min})} x^{m-1} (1-x)^{N-1-m}\, dx \\ & - \frac{N}{N-m-1} \left(\begin{array}{c} N-2 \\ m \end{array} \right) \int_0^{F(a_{min})} x^{m} (1-x)^{N-1-m-1}\, dx \end{aligned}$$

or

$$\Delta(m) = \frac{N}{(N-m)(N-m-1)} \left(\begin{array}{c} N-2 \\ m \end{array} \right) \int_0^{F(a_{min})} x^{m-1} (1-x)^{N-2-m} \{m - Nx\}\, dx$$

The sign of $\Delta(\cdot)$ is determined by the term in brackets which is first positive and then negative. Hence, there is a unique m^* such that $\Delta(m)$ is positive for $m < m^*$, zero at m^* and negative for $m > m^*$. The LHS of (D.7) describes the maximum participation cost that can be supported for a given a_{min}, and we have shown that it is single-peaked in m. Since the function $\frac{1}{m} \cdot F_{N-m:\, N-1}(\cdot)$ is strictly increasing, we have that a_{min} strictly increases in c_p for a given m. Taken together, these imply that choosing an allocation that maximizes the value of solvers' participation cost that can be supported would induce the minimum possible value of a_{min}.

(b) Fix an arbitrary value of solvers' population N and an arbitrary $m \in \{1, \ldots, \overline{m}\}$. Lemma 7 (d) implies that

$$\frac{1}{m} \cdot F_{N-m:\,N-1}(a) - \frac{1}{m} \cdot F_{N+1-m:\,N}(a) > 0, \quad \text{for all } a \in [a_0,\, 1]$$

and for all distributions $F(\cdot)$ that are strictly increasing in their support. Hence, by definition (D.7) for each $\hat{N} < N$ we have that $a_{min}\left(\hat{N}\right) \leq a_{min}(N)$. Lastly, as proved in Theorem 1 the induced threshold ability a_{min} is the same for all solvers, and it does not depend on their ability realization.

(c) By its definition (1.2) a_{min} does not depend on γ and is a function of the order statistics distribution of $F(\cdot)$. □

Proof of Lemma 2. Fix an arbitrary $m \in \{1, \ldots, \overline{m}\}$. We have

$$\Pi_N(m) = \mathbb{E}\left[\sum_{i=1}^{N} \{x_i^*(\mathcal{A}_i;\, m) \text{ is ranked } i\text{th out of } N\} \cdot \mathbb{1}_{\{\text{solver } i \text{ participates}\}}(m)\right]$$

$$= \int_{a_{min}(m)}^{1} \sum_{i=1}^{N} \{\gamma\, a_i + (1-\gamma)\; e_i^*(a_i;\, m)\} \cdot f(a_i)\, da_i$$

$$= N \cdot \int_{a_{min}(m)}^{1} \{\gamma\, a + (1-\gamma)\; e^*(a;\, m)\} \cdot f(a)\, da$$

As shown in Moldovanu et al. (2007) p.357 we have

$$\int_{a_{min}(m)}^{1} \left(\int_{a_{min}(m)}^{a} x \cdot f_{N-m:\,N-1}(x)\; dx\right) \cdot f(a)\, da = \frac{m}{N} \cdot \mathbb{E}[m,\, N;\, a_{min}(m)],$$

where we set $\mathbb{E}[m,\, N;\, a_{min}(m)] := \int_{a_{min}(m)}^{1} x \cdot f_{m:\,N-1}(x)\; dx$. Hence,

$$\begin{aligned}
\Pi_N(m) &= N \cdot \int_{a_{min}(m)}^{1} \{\gamma\, a + (1-\gamma)\; e^*(a;\, m)\} \cdot f(a)\, da \\
&= N \cdot \left(\gamma \int_{a_{min}(m)}^{1} a \cdot f(a)\, da\right) + N\gamma \int_{a_{min}(m)}^{1} (a_{min}(m) - a)\, f(a)\, da \\
&\quad + (1-\gamma)\; N\frac{R}{m} \int_{a_{min}(m)}^{1} \left(\int_{a_{min}(m)}^{a} x\, f_{N-m:\,N-1}(x)\; dx\right) f(a)\, da \\
&= \gamma\, N\, a_{min}(m) \cdot (1 - F(a_{min}(m))) + (1-\gamma)\; R \cdot \mathbb{E}[m,\, N;\, a_{min}(m)]
\end{aligned}$$

which completes the statement. □

Lemma 9. *(a)* $\sum_{i=1}^{N} \mathbb{E}[\mathcal{A}_{i:N}] = N\, \mathbb{E}[\mathcal{A}]$.

(b) If $\mathcal{X}$ *and* $\mathcal{Y}$ *are two random variables such that* $\mathcal{X} \leq_{st} \mathcal{Y}$*, then for any measurable set* $A \subseteq \mathbb{R}$ *we have* $\mathcal{X} \cdot \mathbb{1}_A \leq_{st} \mathcal{Y} \cdot \mathbb{1}_A$.

Proof of Lemma 9. (a) By definition and linearity, $\sum_{i=1}^{N} \mathbb{E}[\mathcal{A}_{i:N}] = \int x \left\{\sum_{i=1}^{N} f_{i:N}(x)\right\} dx$. Let $S(x) := \sum_{i=1}^{N} f_{i:N}(x)$. We have

$$S(x) = \sum_{i=1}^{N} N \begin{pmatrix} N-1 \\ i-1 \end{pmatrix} F^{i-1}(x)\, (1 - F(x))^{N-i}\, f(x)$$

$$= \frac{f(x)}{F(x)} \sum_{i=1}^{N} i \begin{pmatrix} N \\ i \end{pmatrix} F^{i-1}(x) (1 - F(x))^{N-i}$$

In the second equality, we used the fact that the respective sum is simply the expectation of a binomial distribution Binomial $(N, F(x))$, which equals $N \cdot F(x)$. The above implies that $\sum_{i=1}^{N} \mathbb{E}[\mathcal{A}_{i:N}] = \int xS(x)\,dx = N \int x\, f(x)\,dx$.

(b) The statement follows by Theorem 1.C.6 and Theorem 1.C.1 of Shaked and Shanthikumar (2007). □

Lemma 10. *Let f_1 and f_2 be continuous functions on $[a_0, 1]$. Assume that:*

(1) $\int_{a_0}^{1} f_1(x)\,dx \geq 0$

(2) f_1 changes sign from negative to positive at a unique $x_0 \in [a_0, 1]$

(3) f_2 is a (weakly) increasing function

Then, we have that $\int_{a_0}^{1} f_1(x) f_2(x)\,dx \geq 0$.

Proof of Lemma 10. We have that

$$\begin{aligned} \int_{a_0}^{1} f_1(x) f_2(x)\,dx &= \int_{a_0}^{x_0} f_1(x) f_2(x)\,dx + \int_{x_0}^{1} f_1(x) f_2(x)\,dx \\ &\geq f_2(x_0) \int_{a_0}^{x_0} f_1(x)\,dx + f_2(x_0) \int_{x_0}^{1} f_1(x)\,dx \end{aligned}$$

which implies the statement. □

Proof of Theorem 3. Assume that $c_p > 0$ and fix an arbitrary $m \in \{1, \ldots, \overline{m}\}$. To simplify our notation, we assume differentiability and we apply Leibniz's rule of "differentiation under the integral sign" treating m as a continuous variable (a discrete analog holds as well):

$$\begin{aligned} \frac{d\Pi_N}{dm}(m) &= N \cdot \frac{d}{dm} \int_{a_{min}(m)}^{1} \left\{ \gamma a + (1-\gamma)\, e^*(a;\, m) \right\} \cdot f(a)\,da \\ &= N(1-\gamma) \int_{a_{min}(m)}^{1} \frac{\partial e^*}{\partial m}(a;\, m)\, f(a)\,da - N \frac{\partial a_{min}}{\partial m}(m) \cdot f(a_{min}(m)) \cdot \Big(\gamma\, a_{min}(m) + (1-\gamma)\, \cancelto{0}{e^*(a_{min}(m);\, m)} \Big) \\ &= \underbrace{N \cdot (1-\gamma) \int_{a_{min}(m)}^{1} \frac{\partial e^*}{\partial m}(a;\, m) \cdot f(a)\,da}_{\text{effort effect}} \underbrace{- N \cdot \gamma\, a_{min}(m) \cdot \frac{\partial a_{min}}{\partial m}(m)}_{\text{participation effect}} \end{aligned}$$

(a) If contest specialization is zero ($\gamma = 0$), then the aforementioned "participation effect" disappears and by Lemma 2 we have that

$$\begin{aligned} \Pi_N(m) &= R \cdot \mathbb{E}[m,\, N;\, a_{min}(m)] \\ &= R \cdot \mathbb{E}[\mathcal{A}_{N-m:\,N} |\ \mathcal{A} \geq a_{min}(m)] \end{aligned}$$

Due to stochastic dominance, Lemma 7(a) and (b) imply that $\mathbb{E}[\mathcal{A}_{N-1:\,N}] \geq \mathbb{E}[\mathcal{A}_{N-m\,N}]$ for all $m \in \{1, \ldots, N-1\}$. Hence, by Lemma 9(b) we have that $\Pi_N(1) \geq \Pi_N(m)$ for all $m \in \{1, \ldots, N-1\}$, or that WTA is optimal when $\gamma = 0$ and $c_p > 0$. Further, note that the statement follows also by substituting $k := N$ and $\gamma := 0$ into Theorem 5, due to Lemma 9(a) that relates the expected value of a random variable to the total sum of the expected order statistics.

(b) Theorem 3(a) shows that the "effort effect" is zero at $m = 1$. Further, $a_{min}(m) > 0$ and Corollary 1(a) shows that there exists a unique $m_0^* \in \{1, \ldots, \overline{m}\}$ such that $a_{min}(s) > a_{min}(m_0^*)$

for all $s \in \{1, \ldots, m_0^* - 1\}$, and $a_{min}(m_0^*) < a_{min}(s)$ for all $s \in \{m_0^* + 1, \ldots, N-1\}$. For $\gamma \in (\hat{\gamma}, 1]$ participating solvers exert zero effort and in that case it is optimal for the firm to minimize the induced a_{min}, that is to allocate m_0^* equal awards to the top m_0^* ranking positions.

For $\gamma \in (0, \hat{\gamma}]$ we consider two cases:

Case 1 ($m > m_0^*$): $\frac{d\Pi_N}{dm}(m) < 0$, hence $\Pi_N(m_0^*) > \Pi_N(m)$.

Case 2 ($m \leq m_0^*$): $\frac{d\Pi_N}{dm}(m)$ is single crossing from positive to negative, hence a unique $\hat{m} \in \{1, \ldots, m_0^*\}$ exists such that $\Pi_N(\hat{m}) > \Pi_N(m)$. That is, WTA is not always optimal for $\gamma \in (0, \hat{\gamma}]$. □

Proof of Theorem 4. Consider the budget-to-participation cost ratio $r := \frac{R}{c_p}$. We show below the dependence of m^* on r holding all else parameters constant.

We have shown in Theorem 1 that solvers' expected utility is strictly increasing in ability. By its definition (1.2), the ability threshold a_{min} is increasing in r, i.e. for each $\hat{r} < r$ we will have that $\hat{a}_{min}(\hat{r}) < a_{min}(r)$. We first show that $m_0^* = m_0^*(r)$ is weakly decreasing in r. To show that $m_0^*(r, a_{min}) = \arg\max_{1 \leq j \leq \overline{m}} \left\{ \frac{F_{N-j:N-1}(a_{min})}{j} \right\}$ is (weakly) decreasing in r, it suffices to show that $m_0^*(r, a_{min}(r))$ is (weakly) decreasing in a_{min} for a fixed r. Corollary 1 shows that for a given a_{min}, the ratio $A_j(a_{min}) := \frac{F_{N-j:N}(a_{min})}{j}$ has a unique maximum wrt j. Further, the function $h(a_{min}) := \max_{1 \leq j \leq \overline{m}} A_j(a_{min})$ is increasing in a_{min}. Hence, $\arg\max_{1 \leq j \leq \overline{m}} A_j(\hat{a}_{min}) \geq \arg\max_{1 \leq j \leq \overline{m}} A_j(a_{min})$, which implies that $m_0^*(\hat{r}, \hat{a}_{min}(\hat{r})) \geq m_0^*(r, a_{min}(r))$. Finally, by Theorem 3(b) we have that by definition m^* can not increase in r.

(a) and (c) Since $m^* = m^*(r)$ is weakly decreasing in r, for a fixed c_p we have that $m^* = m^*(R)$ is weakly increasing in R. Further, for a fixed R we have that $m^* = m^*(c_p)$ is weakly decreasing in c_p.

(b) Seeker's objective (1.5) implies that for an arbitrary $\gamma \in (\hat{\gamma}, 1]$: $m^* = m_0^*$ is the optimal number of awards, which does not depend on γ (see (Corollary 1)(c)). Also, for $\gamma = 0$ we have shown in Theorem 3(a) that WTA is optimal. For an arbitrary $\gamma \in (0, \hat{\gamma}]$ the first terms in seeker's objective (1.5) are maximized at $m_0^* \geq 1$ (these terms are strictly increasing in γ), and the last term (which is strictly decreasing in γ) is maximized at $m^* = 1$. Hence, as γ increases (i.e. as the weight on solvers' effort that determines the rank of their performance decreases) the $m^* = \arg\max_{1 \leq m \leq \overline{m}} \Pi_N(m; \gamma)$ cannot decrease.

(d) We first prove that $m_0^* = m_0^*(N)$ is weakly increasing in N. We have shown that for a fixed a_{min}, $m^*(c_p, a_{min})$ is (weakly) decreasing in c_p. We further have that all else equal, c_p increases in a_{min}, and Corollary 1(b) shows that a_{min} is weakly increasing in solvers' population N. Hence, $m_0^* = m_0^*(N)$ is weakly increasing in N. Finally, by Theorem 3(b) we have that by definition m^* can not decrease in N. □

Proof of Theorem 5. Fix an arbitrary $m \in \{1, \ldots, \overline{m}\}$ and an arbitrary $k \in \{1, \ldots, N\}$. We have

$$\begin{aligned}
\Pi_k(m) &= \mathbb{E}\left[\sum_{i=1}^{k} w_i \cdot \{x_i^*(\mathcal{A}_i; m) \text{ is ranked } i\text{th out of } N\} \cdot \mathbb{1}_{\{\text{solver } i \text{ participates}\}}(m) \right] \\
&= \gamma\, \mathbb{E}\left[\sum_{i=1}^{k} w_i \cdot \{\mathcal{A}_i \text{ is ranked } i\text{th out of } N\} \,\middle|\, \mathcal{A}_i \geq a_{min}(m) \right]
\end{aligned}$$

$$+ (1-\gamma)\ \mathbb{E}\left[\sum_{i=1}^{k} w_i \cdot \{e_i^*(\mathcal{A}_i;\, m)\ \text{ is ranked } i\text{th out of } N\}\,\middle|\, \mathcal{A}_i \geq a_{min}(m)\right]$$

$$= \gamma \sum_{i=1}^{k} w_i \cdot \int_{a_{min}(m)}^{1} a\, dF_{N-i:\,N}(a) + (1-\gamma) \sum_{i=1}^{k} w_i \cdot \int_{a_{min}(m)}^{1} e^*(a)\ dF_{N-i:\,N}(a)$$

The rest follows by Theorem 16 where we solve the general version of seeker's problem. □

D.4 The optimal allocation of prizes in contests with endogenous participation and unobservable effort that maximizes the best k participating performers

In this section we solve seeker's general problem. Assume that the seeker determines the *number* of rewards having positive value and the entire *distribution* of her total budget R among the different rewards for each rank, in order to maximize the expected value of a weighted combination of the top $k \in \{1, 2, \ldots, N-1\}$ *participating* solvers (ranked by their equilibrium performance in relative order), conditional that at least k solvers participate. If less or equal than k solvers participate (an event that the seeker wishes to happen with sufficiently low probability, if not with *zero* probability), then we set seeker's objective to zero. It is easy to see that Proposition 2 can be extended to show that having a budget

$$k\, c_p < R < N\, c_p$$

is an necessary and sufficient condition to guarantee that the event $\{\mathcal{N} \leq k-1\} \cup \{\mathcal{N} = N\}$ has zero probability. Alternatively, since the number of participating solvers follows Binomial distribution, we can use Chernoff bounds and show that the positive probability $\mathbb{P}[\mathcal{N} \leq k-1]$ can be made sufficiently small so that the seeker can ignore it. We focus on weakly monotone allocations of non-negative rewards and we impose the budget constraint $\sum_{j=1}^{N} R_j \leq R$. That is, if no solvers participate, the seeker keeps her budget; if one solver participates he exerts zero effort and gets the top reward R_1, and so on and so forth for all other special cases.

Seeker's problem is given by:

$$\begin{aligned}
&\max_{R_1,\ldots,R_N} \hat{\Pi}\left((R_j)_{j=1}^{N}\right) \\
&\text{s.t.} \quad u(a_i,\, e_i^*;\, \tilde{p}) \geq c_p, \quad \text{for all } i \text{ that enter} && \text{(IR)} \\
&\qquad e^* = \arg\max_{e \geq 0}\ u(a,\, e;\, \tilde{p}) \quad \text{for all } a && \text{(IC)} \\
&\qquad \sum_{j=1}^{N} R_j \leq R && \text{(Budget constraint)} \\
&\qquad R_j \geq R_{j+1}, \quad j = 1, \ldots, N-1 && \text{(Monotonicity)} \\
&\qquad \tilde{p} = p^* && \text{(SCE)}
\end{aligned} \tag{D.14}$$

where

$$\hat{\Pi}\left((R_j)_{j=1}^N\right) := \mathbb{E}\left[\sum_{i=1}^{k} w_i \left\{\gamma\, \mathcal{A}_{N-i:\,N} + (1-\gamma)\; e^*\left(\mathcal{A}_{N-i:\,N};\, (R_j)_{j=1}^N\right)\right\} \cdot \mathbb{1}_{\{i \text{ enters}\}}\left((R_j)_{j=1}^N\right) \middle|\, \mathcal{N} \geq k\right]$$

for exogenous weights $w_1 > w_2 > \ldots > w_k$ and

$$u_i\left(a_i,\, e_i^*;\, \tilde{p}\right) = \sum_{j=1}^{N} R_j \cdot \mathbb{P}\left[x_i^* \text{ ranked } j\text{th out of } N;\, \tilde{p}\right] - \frac{e_i^*}{a_i} - c_p$$

Note that the reward allocation chosen by the seeker determines the participation decision of the solvers, as well as their effort action, but the abilities of the solvers are not known by the seeker and we denote them with random variables. The seker

The optimization (D.14) is a combinatorial problem over the *number* of non-zero rewards to allocate as well as their *size*, for all possible combinations that satisfy the constraints above. To solve it, we first characterize solvers' best response to a fixed reward allocation chosen by the seeker. We focus on pure symmetric self-confirming equilibria (SCE). The result follows by the proofs of Theorem 1 and Theorem 2 and we omit its proof.

Theorem 15 (Self-confirming equilibrium - general case)**.** *(a) There exists a unique pure symmetric equilibrium characterized by a couple* $(p^*,\, a_{min}) \in [0,\, 1] \times [a_0,\, 1]$ *that solves*

$$\left.\begin{array}{c} \sum_{j=1}^{N} \left(R_j - R_{j+1}\right) \cdot F_{N-j:\,N-1}\left(a_{min}\left(p^*\right)\right) = c_p \\ p^* = 1 - F\left(a_{min}\left(p^*\right)\right) \end{array}\right\} \tag{D.15}$$

such that $\mathcal{N} \sim$Binomial$(N,\, 1 - F\left(a_{min}\right))$ *solvers participate in equilibrium according to beliefs* p^* *on the fraction of participants.*

(b) There exists $\tilde{\gamma} \in [0,\, 1]$ *that depends on the allocation* $(R_j)_{j=1}^N$ *such that for any fixed contest specialization* $\gamma \in [\tilde{\gamma},\, 1]$*:* $e^*\left(a;\, \gamma\right) = 0$ *for all* $a \in [a_0,\, 1]$*. When* $\gamma \in [0,\, \tilde{\gamma})$ *a solver with ability* a *exerts equilibrium effort*

$$e^*\left(a;\, \gamma,\, (R_j)_{j=1}^N\right) = \begin{cases} \frac{\gamma\left(a_{min} - a\right)}{1-\gamma} + \sum_{j=1}^{N}\left(R_j - R_{j+1}\right) \int_{a_{min}}^{a} x\, dF_{N-j:\,N-1}\left(x\right), & a \geq a_{min} \\ 0, & a < a_{min} \end{cases} \tag{D.16}$$

It is crucial to note that the allocation of rewards affects the functional form of solvers' effort, the lower limit of the integral in (D.16), as well as the participation decision of the solvers. These correspond to the "effort", "screening" and "participation" effects respectively. As we show below, all these three effects *combined* determine the optimal solution to seeker's objective function in (D.14). In addition, the first term of (D.16) is decreasing in ability, whereas the second term is increasing in ability. Depending on which effects dominates, the equilibrium effort is in general single-peaked.

To guarantee that at least k solvers participate with non-trivial probability $p^* \in (0,\, 1)$ and exert strictly positive effort we assume that the following generalized version of Budget Condition holds:

$$\max\left\{\frac{1}{a_0},\, k \cdot c_p\right\} < R < N\, c_p, \qquad \text{(Generalized Budget Condition)}$$

for $k \in \{1, \ldots,\, N-1\}$. If the firm cares about the case $k = N$, we assume the Budget Condition

to make our setting non-trivial. This implies that $\tilde{\gamma} = 1$, hence all participating solvers exert strictly positive effort in equilibrium for all contest specializations $\gamma \in [0,\, 1)$. Trivially, if $\gamma = 1$ then solver effort has no impact on his performance rankings and all participating solvers exert zero effort upon entry.

Armed with solvers' best response to a given reward allocation chosen by the seeker, we provide a characterization of the solution to seeker's combinatorial problem (D.14). The following Theorem shows that the structure of the optimal allocation has the MW format for innovation contests with unobservable effort when the seeker cares about the *total* performance of the *participants*, or for purely ability-based innovation contests and irrespective on the objective of the seeker. Further, we provide a tight upper bound m_0^* on the optimal number of awards set by the seeker, due to the endogenous participation actions of the solvers.

Theorem 16 (Top k participating performers - general case)**.** *Seeker's objective is written as*

$$\hat{\Pi}\left((R_j)_{j=1}^N\right) = \gamma \sum_{i=1}^{k} w_i \cdot N \int_{a_{min}\left((R_j)_{j=1}^N\right)}^{1} a\, \{F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a)\}\, dF(a)$$
$$+ (1-\gamma) \sum_{i=1}^{k} w_i \cdot N \int_{a_{min}\left((R_j)_{j=1}^N\right)}^{1} e^*\left(a;\, (R_j)_{j=1}^N\right) \{F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a)\}\, dF(a)$$

Define $(m_0^*,\, a_{min}^*) \in \{1, \ldots, \overline{m}\} \times [a_0,\, 1]$ *as the unique solution of the system*

$$\left.\begin{array}{c} m_0^* = \arg\max_{1 \le j \le \overline{m}} \left\{ \frac{F_{N-j:\,N-1}\left(a_{min}^*\right)}{j} \right\} \\ \frac{R}{m_0^*} \cdot F_{N-m_0^*:\,N-1}\left(a_{min}^*\right) = c_p \end{array}\right\} \tag{D.17}$$

The optimal allocation to (D.14) *has at most* m_0^* *non-zero awards characterized as follows*

$$\left(R_j^*\right)_{j=1}^{m_0^*} = \begin{cases} \left(\underbrace{\frac{R}{m_0^*},\, \frac{R}{m_0^*}, \ldots,\, \frac{R}{m_0^*}}_{m_0^* - \textit{equal rewards}},\, 0,\, 0, \ldots,\, 0 \right), & \gamma = 1,\ \textit{and}\ 1 \le k \le N \\ (R,\, 0,\, 0, \ldots,\, 0), & \gamma \in [0,\, 1),\ \textit{and}\ 1 \le k \le N-1 \\ (R,\, 0,\, 0, \ldots,\, 0), & \gamma = 0,\ \textit{and}\ k = N \\ \left(\underbrace{R_1^*,\, R_2^*, \ldots,\, R_{m^*}^*}_{m^* \le m_0^*\ \textit{rewards}},\, 0, \ldots,\, 0 \right), & \gamma \in (0,\, 1),\ \textit{and}\ k = N \end{cases} \tag{D.18}$$

Proof of Theorem 16. From Lemma 1 we have that awarding more than $\overline{m}$ distinct prizes induces no solvers to participate. Hence, the optimal number of rewards to set is strictly less than $\overline{m}$. We consider four special cases. In all cases below we assume a strictly positive participation cost $c_p > 0$ and seeker's budget satisfies the Generalized Budget Condition.

Case 1: $\gamma = 1$ and $k \in \{1, \ldots, N\}$. Solver's objective is given by

$$\hat{\Pi}\left((R_j)_{j=1}^{\overline{m}}\right) = \sum_{i=0}^{k-1} w_i \cdot \int_{a_{min}\left((R_j)_{j=1}^{\overline{m}}\right)}^{1} a\, dF_{N-i:\,N}(a)$$

which is maximized at the allocation $(R_j)_{j=1}^{\overline{m}}$ that *minimizes* solvers' unique participation threshold a_{min}. Recall that the LHS of the (IR) constraint that defines a_{min} is *strictly* increasing in a_{min}, and hence it is invertible. To minimize a_{min} we can equivalently maximize the maximum participation cost that can be supported for a given award allocation, over all monotone allocations. The latter is the solution to

$$\begin{aligned} \max_{R_1,\ldots,R_{\overline{m}}} & \sum_{j=1}^{\overline{m}-1} (R_j - R_{j+1}) \cdot F_{N-j:\,N-1}(a_{min}) \\ \text{such that} \quad & R_j \geq R_{j+1}, \quad j = 1,\ldots,\overline{m}-1 \\ & \sum_{j=0}^{\overline{m}-1} R_j = R \end{aligned} \tag{D.19}$$

Note that the budget constraint is binding. Define $\delta_j := j \cdot (R_j - R_{j+1})$ for $j = 1,\ldots,\overline{m}$ and observe that the linear program (D.19) can be written as

$$\begin{aligned} \max_{\delta_1,\ldots,\delta_{\overline{m}}} & \sum_{j=1}^{\overline{m}} \delta_j \cdot \frac{F_{N-j:\,N-1}(a_{min})}{j} \\ \text{such that} \quad & \delta_j \geq 0, \quad j = 1,\ldots,\overline{m} \\ & \sum_{j=1}^{\overline{m}} \delta_j = R \end{aligned}$$

Corollary 1 shows that for a given a_{min} the function $g(m) := \frac{1}{m} \cdot F_{N-m:\,N-1}(a_{min})$ is unimodal and is maximized at a unique m_0^* defined as:

$$m_0^* := \arg \max_{1 \leq j \leq \overline{m}} \left\{ \frac{F_{N-j:N-1}(a_{min})}{j} \right\}$$

The latter definition combined with the (IR) condition (D.15) implies that the optimal pair $(m_0^*,\, a_{min}^*)$ is the unique solution to the system (D.17) for $\gamma = 1$.

Case 2: $\gamma = 0$ and $k = N$. In this case, the seeker cares about the total effort (which equals performance) of the participating solvers. By Theorem 15 and $\gamma = 0$ we have that

$$e^*\left(a;\, (R_j)_{j=1}^N\right) = \sum_{j=1}^{N} (R_j - R_{j+1}) \int_{a_0}^{a} x\, f_{N-j:\,N-1}(x)\, dx$$

Observe that the term corresponding to $j := N$ is negative, so we have that $R_N^* = 0$. It suffices to show that at optimality all differences in rewards from the 2nd position till N are zero, hence $R_j^* = R_N^* = 0$ for all $j \geq 2$. Assume that $R_1 - R_2 = \delta R$ and $R_2 = (1-\delta)\, R$, for $\delta \in [0,\, 1]$. Seeker's objective becomes

$$\max_{\delta \in [0,\,1]} \int_{a_{min}(\delta)}^{1} e^*(a;\, \delta)\, dF(a)$$

Due to the boundary condition $e^*(a_{min}(\delta);\, \delta) = 0$ we have that

$$\frac{d\Pi}{dm}(\delta) = \int_{a_{min}(\delta)}^{1} \frac{de^*}{d\delta}(a;\, \delta)\, dF(a) - \frac{\partial a_{min}}{\partial \delta}(\delta) \cdot f(a_{min}(\delta)) \cdot \cancelto{0}{e^*(a_{min}(\delta);\, \delta)}$$

which implies that seeker's problem becomes

$$\max_{\delta\in[0,\,1]} \delta R \int_{a_{min}}^{1} \left[\int_{a_{min}}^{a} x\ \left(f_{N-1:\,N-1}\left(x\right) - f_{N-2:\,N-1}\left(x\right)\right)\ dx\right]\ dF\left(a\right)$$

Similar to the proof of Moldovanu and Sela (2001), showing that the integral sign is strictly positive for all F shows that the optimal value of δ is $\delta^* = 1$. Indeed, by Lemma 1.6 of Moldovanu and Sela (2001) we have that $f_{N-1:\,N-1}\left(x\right) - f_{N-2:\,N-1}\left(x\right)$ changes sign from negative to positive at a unique $x_0 \in [a_0,\,1]$. Hence, seeker's objective $\hat{\Pi}\left(\left(R_j\right)_{j=1}^{N}\right)$ is maximized at a WTA allocation for $\gamma = 0$ and $k = N$.

Case 3: $\gamma \in [0,\,1)$ and $1 \leq k \leq N-1$. In this case, the seeker cares about the total performance of the participating solvers. We assume that seeker's budget satisfies the Budget Condition. Due to the threshold participation strategy of the solvers and Lemma 8, being ranked kth highest among the top subset of a population of random size is equal to being ranked kth highest out of the entire population, if and only if, the top subset has size greater or equal to k with probability one. Hence, we can equivalently solve seeker's problem when all solvers participate. Corollary 2 shows that a WTA allocation is optimal in this case.

Case 4: $\gamma \in (0,\,1)$ and $k = N$. Arguing similarly to Theorem 3(b) we have that it is not optimal to allocate more than m_0^* awards since it would hurt both the "effort effect" as well as the "participation effect". In particular, the "Case 1" above shows that to maximize the "participation effect" we should equally spread the budget among the top m_0^* solvers, allocating zero awards to any lower ranked participating solvers, whereas "Case 2" shows that the "effort effect" is maximized by allocating all the budget to the top. For intermediate values of $\gamma \in (0,\,1)$ checking all combinations of $m \in \{1,\ldots,\,m_0^*\}$ awards rewarded to the top m solvers leads to the allocation that maximizes $\hat{\Pi}\left(\cdot\right)$. □

Corollary 2 (Top k performers with exogenous participation)**.** *Let N be the solver population size which we allow to be finite or infinite, and assume that all N solvers participate with certainty. Then, the WTA allocation is optimal for all $\gamma \in [0,\,1]$ and $1 \leq k \leq N$.*

Proof of Corollary 2. Fix an arbitrary $k \in \{1,\ldots,\,N\}$. We have that N solvers participate w.p.1, if and only if, $R \geq N \cdot c_p$. When $N \to +\infty$, the latter condition is satisfied for finite R, if and only if, $c_p = 0$. Implying these conditions, Archak and Sundararajan (2009) show that when WTA is optimal for $\gamma = 0$ and $1 \leq k \leq N$. This also holds for $\gamma \in [0,\,1]$ and $1 \leq k \leq N$ since the observability of effort does not affect seeker's objective when solver participation is exogenously guaranteed.

Assume that $R \geq N \cdot c_p$ and $N < \infty$. From Theorem 16 with $a_{min} := a_0$ and $w_i = 1$ for all i, seeker's objective is written as:

$$\begin{aligned}\hat{\Pi}\left(\left(R_j\right)_{j=1}^{N}\right) = \gamma \sum_{i=1}^{k} w_i \cdot N \int_{a_0}^{1} w_i \cdot a\ \left\{F_{N-i-1:\,N}\left(a\right) - F_{N-i:\,N}\left(a\right)\right\}\ dF\left(a\right)\\ + \left(1-\gamma\right) \sum_{i=1}^{k} w_i \cdot N \int_{a_0}^{1} e^*\left(a;\,\left(R_j\right)_{j=1}^{N}\right)\ \left\{F_{N-i-1:\,N}\left(a\right) - F_{N-i:\,N}\left(a\right)\right\}\ dF\left(a\right)\end{aligned}$$

Note that only the second term is affected by the reward allocation $\left(R_j\right)_{j=1}^{N}$. Moldovanu and

Sela (2001) show that $N \int_{a_0}^{1} e^* \left(a;\, (R_j)_{j=1}^N\right) dF(a) \geq 0$ is maximized at a WTA allocation that places all the budget at the top. To show that this also holds for $\hat{\Pi}\left((R_j)_{j=1}^N\right)$ it suffices to show that the optimal value of the second reward $R_2^* = 0$ (since we focus on weakly monotone allocations).

Setting $\gamma = 0$ into Theorem 15 we have that

$$e^*\left(a;\, (R_j)_{j=1}^N\right) = \sum_{j=1}^{N} (R_j - R_{j+1}) \int_{a_0}^{a} x\, f_{N-j:\,N-1}(x)\, dx$$

Observe that the term corresponding to $j := N$ is negative, so we have that $R_N^* = 0$. It suffices to show that all differences in rewards from the 2nd position till N are zero, hence $R_j^* = R_N^* = 0$ for all $j \geq 2$. Assume that $R_1 - R_2 = \delta R$ and $R_2 = (1-\delta) R$, for $\delta \in [0,\, 1]$.

Fix an arbitrary index $i \in \{1, \ldots,\, k\}$. Seeker's objective becomes

$$\max_{\delta \in [0,\,1]} \int_{a_0}^{1} e^*(a;\, \delta)\, \{F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a)\}\, dF(a)$$

or equivalently

$$\max_{\delta \in [0,\,1]} \delta R \int_{a_0}^{1} \left[\int_{a_0}^{a} x\, (f_{N-1:\,N-1}(x) - f_{N-2:\,N-1}(x))\, dx\right] \{F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a)\}\, dF(a)$$

Similar to the proof of Moldovanu and Sela (2001), it suffices to show that the integral sign is strictly positive for all F.

By Lemma 1.6 of Moldovanu and Sela (2001) we have that $f_{N-1:\,N-1}(x) - f_{N-2:\,N-1}(x)$ changes sign from negative to positive at a unique $x_0 \in [a_0,\, 1]$. Also, by B.3 we have that the function

$$F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a) = \left(\begin{array}{c} N-1 \\ j-1 \end{array}\right) F(a)^{N-j} (1 - F(a))^{j-1}$$

is strictly increasing in a, for all $a \in [a_0,\, 1]$.

Applying Lemma 10 we immediately have that

$$\int_{a_0}^{1} \left[\int_{a_0}^{a} x\, (f_{N-1:\,N-1}(x) - f_{N-2:\,N-1}(x))\, dx\right] \{F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a)\}\, dF(a) > 0,$$

or that a WTA allocation maximizes

$$N \int_{a_0}^{1} e^*\left(a;\, (R_j)_{j=1}^N\right) \{F_{N-i-1:\,N}(a) - F_{N-i:\,N}(a)\}\, dF(a)$$

Hence, seeker's objective $\hat{\Pi}\left((R_j)_{j=1}^N\right)$ is maximized at a WTA allocation for all $\gamma \in [0,\, 1]$ and $1 \leq k \leq N$. □

Corollary 2 generalizes the classic WTA result of Moldovanu and Sela (2001) which was established for the seeker objective of maximizing the total sum of the efforts which are observable by the seeker ($k := N$ and $\gamma := 0$), the solver population size is finite ($N < \infty$), and all solvers participate with certainty. Further, Corollary 2 shows that the result of Archak and Sundararajan (2009) holds also for finite population size, unobservable effort and no participation uncertainty.

This page is intentionally left blank

Appendix $\mathcal{E}$

Proofs of Chapter 2

E.1 Summary of notation used

Firm:

- k: number of distinct priority classes to form.
- N_j: size of priority class j, where $j \in \{1, \ldots, k\}$. We denote by $\boldsymbol{N} := (N_1, \ldots, N_k)$ the partition chosen by the firm and require $\sum_{j=1}^{k} N_j = N$.

Agents:

- $\mathcal{N}_j$: random variable representing the *ex ante* unknown number of participating agents in priority class j, where $j \in \{1, \ldots, k\}$. The support of $\mathcal{N}_j$ is $\{0, 1, \ldots, N_j\}$.
- $\boldsymbol{\mathcal{N}} := (\mathcal{N}_1, \ldots, \mathcal{N}_k)$ is the random vector of the *ex ante* unknown number of participating agents in each priority class. Similarly, let $\mathcal{N} := \sum_{j=1}^{k} \mathcal{N}_j$ be the random variable with support $\{0, 1, \ldots, N\}$ of the *ex ante* unknown total number of participating agents.
- $\boldsymbol{n} := (n_1, \ldots, n_k)$ is the realization of $\boldsymbol{\mathcal{N}}$, *ex post* the participation decision of the strategic agents. Similarly, let $n := \sum_{j=1}^{k} n_j$ be the *ex post* realized total number of participating agents.

Other parameters:

- $\lambda = \lambda(\mathcal{N})$: (endogenous) mean arrival rate arriving into the system in total. Conditional that $\{\mathcal{N} = n\}$ agents participate, $\lambda(n)$ is the mean arrival rate that solves (2.1).
- λ_j; ρ_j, $\hat{\rho}_j$: realized demand rate allocated to class j according to FRFtS (where $j \in \{1, \ldots, k\}$); realized and expected utilization of class j, respectively.
- Λ, V, c: (exogenous) mean customer demand for service, customers' valuation of service, and cost per unit of time waiting, respectively.
- V_f, w: (exogenous) revenue per sale for the firm, and (exogenous) hourly wage paid to agents when utilized respectively.

- N: (exogenous) size of agents' population.
- c_p: (exogenous) agents' participation cost.
- $F(\cdot)$, $f(\cdot)$: (exogenous) CDF and PDF of agents' ability distribution, which is assumed common knowledge and strictly increasing in its support $[0,1]$. We let $\mathcal{A}_i$ denote the ability of an agent i which is a random variable for the firm, whose realization $\{\mathcal{A}_i = a_i\}$ is privately known to agent i.
- $\tilde{\beta}$: (exogenous) agents' *ex ante* distribution of the (symmetric) participation actions of the others. In particular, the agents *ex ante* believe that their peers participate with (symmetric) probability $\tilde{p}$.
- β^*: (endogenous) agents' equilibrium distribution of the participation actions of the others. In equilibrium, the agents participate with (symmetric) probability p^*.

E.2 The $M/M/\mathcal{N}$ model with Ranked Servers

In this section, we analyze the queuing dynamics of our setting described in §2.3. Following the sequence of events in §2.3, first $\{\mathcal{N} = n\}$ servers enter the system (Step 2) and then congestion-sensitive demand $\lambda(n)$ is realized (Step 3), where $\lambda(n)$ is the unique solution to (2.1) and it scales appropriately so that $\lambda(n) < n$, for every realization $n \in \{1, \ldots, N\}$ of the number of participating agents with $\lambda(0) = 0$. Since demand arrivals are Poisson and service times are exponentially distributed, we refer to this setting as the $M/M/\mathcal{N}$ model with ranked servers in k priority classes.

First, we examine the $M/M/\mathcal{N}$ model *without* server priorities. Conditional that $\{\mathcal{N} = n\}$ servers enter the system, it is known from the classic $M/M/n$ model with $\mu = 1$ that, if $\mathcal{C}(t)$ denotes the number of customers in the system at time t, the following steady state distribution exists

$$\hat{\pi}(s;\, n) := \lim_{t\to\infty} \mathbb{P}\left[\mathcal{C}(t) = s\right] = \begin{cases} \frac{1}{s!}\lambda(n)^s\, \hat{\pi}(0;\, n), & 0 \le s \le n \\ \frac{1}{n!}\lambda(n)^n \left(\frac{\lambda(n)}{n}\right)^{s-n} \hat{\pi}(0;\, n), & s \ge n \end{cases} \tag{E.1}$$

where $\hat{\pi}(0;\, n) := \left(\sum_{r=0}^{n-1} \frac{\lambda^r}{r!} + \frac{\lambda^n}{n!\left(1-\frac{\lambda}{n}\right)}\right)^{-1}$. Next, we define a relevant performance metric for the expected amount of time an individual agent is busy in an $M/M/\mathcal{N}$ model, averaged over the number of participating agents.

Definition 2 (Expected utilization)**.** Let $\mathcal{N}$ be a non-negative discrete random variable. The expected long-run fraction of the time that a participating agent is busy in the $M/M/\mathcal{N}$ model without server priorities is

$$\hat{\rho} := \mathbb{E}_{\mathcal{N}}\left[\sum_{s=1}^{\mathcal{N}-1} \frac{s}{\mathcal{N}}\, \hat{\pi}(s;\, \mathcal{N}) + \sum_{s=\mathcal{N}}^{\infty} \hat{\pi}(s;\, \mathcal{N}) \,\middle|\, \mathcal{N} \ge 1\right] \tag{E.2}$$

We refer to $\hat{\rho}$ as the *expected utilization* of a participating agent.

The number of participating servers is only known *ex post* agents' participation decision. Hence, the above definition captures the *a priori* utilization value of an individual agent when no priority classes are formed. The above definition makes no assumption on the distribution of the number of participating agents. As we prove in §2.4, the Binomial distribution arises in equilibrium by the strategic participation choices of the agents. By conditioning on the number of participating agents and using (E.1), we show next that agents' expected utilization (E.2) under congestion-sensitive demand increases in participation and can be simplified as follows.

Lemma 11. *Conditional on* $\{\mathcal{N} = n\}$ *for* $n \in \{1, \ldots, N\}$*, let* $\lambda(n)$ *be the unique solution to* (2.1) *with* $\lambda(n) < n$ *and* $\lambda(0) = 0$*.*

(a) The expected utilization of a participating agent in the $M/M/\mathcal{N}$ *model is equal to*

$$\hat{\rho} = \mathbb{E}_{\mathcal{N}}\left[\frac{\lambda(\mathcal{N})}{\mathcal{N}}\middle|\, \mathcal{N} \geq 1\right]$$

(b) If $\mathcal{N} \sim Bin(N, p)$ *then the expected utilization* $\hat{\rho}$ *is strictly increasing in agents' participation probability* p*.*

Proof of Lemma 11. Fix a realization $n \in \{0, \ldots, N\}$ of the number of participating agents $\mathcal{N}$. Recall that congestion-sensitive demand with customer expected utility (2.1) implies that the equilibrium demand rate scales appropriately so that $\lambda(n) < n$, and $\lambda(0) = 0$. If $n = 0$, then $\hat{\rho} := 0$ by Definition 2. Using the steady state distribution (E.1) of the $M/M/n$ we get

$$\begin{aligned}
\sum_{s=1}^{n} \frac{s}{n}\hat{\pi}(s;\, n) + \sum_{s=n+1}^{\infty} \hat{\pi}(s;\, n) &= \hat{\pi}(0;\, n) \cdot \left\{\frac{1}{n} \cdot \sum_{s=1}^{n} \frac{\lambda(n)^s}{(s-1)!} + \frac{\lambda(n)^n}{n!} \cdot \sum_{s=n+1}^{\infty} \left(\frac{\lambda(n)}{n}\right)^{s-n}\right\} \\
&= \hat{\pi}(0;\, n) \cdot \left\{\frac{1}{n} \cdot \sum_{r=0}^{n-1} \frac{\lambda(n)^r}{r!} + \frac{\lambda(n)^n}{n!} \cdot \sum_{r=1}^{\infty} \left(\frac{\lambda(n)}{n}\right)^{r}\right\} \\
&= \hat{\pi}(0;\, n) \cdot \left\{\frac{1}{n} \cdot \sum_{r=0}^{n-1} \frac{\lambda(n)^r}{r!} + \frac{\lambda(n)^n}{n!} \cdot \left(\frac{\frac{\lambda(n)}{n}}{1 - \frac{\lambda(n)}{n}}\right)\right\} \\
&= \frac{\lambda(n)}{n},
\end{aligned} \tag{E.3}$$

The last step follows by the definition of $\hat{\pi}(0;\, n)$. Conditioning on the number of participating agents we have

$$\begin{aligned}
\hat{\rho} &= \mathbb{E}_{\mathcal{N}}\left[\sum_{s=1}^{n-1} \frac{s}{n}\hat{\pi}(s;\, n) + \sum_{s=n}^{\infty} \hat{\pi}(s;\, n)\middle|\, \mathcal{N} = n \geq 1\right] = \sum_{n=1}^{N} \frac{\lambda(n)}{n} \cdot \mathbb{P}[\mathcal{N} = n] \\
&= \mathbb{E}_{\mathcal{N}}\left[\frac{\lambda(\mathcal{N})}{\mathcal{N}}\middle|\, \mathcal{N} \geq 1\right]
\end{aligned}$$

Further, Lemma 9(a) of Taylor (2016) shows that the ratio $\frac{\lambda(n)}{n}$ strictly increases in n, for each $n \in \{1, \ldots, N\}$. Then, a coupling argument implies that if $\mathcal{N}_1 \sim \text{Bin}(N, p_1)$ and $\mathcal{N}_2 \sim \text{Bin}(N, p_2)$ with $p_1 < p_2$, then $\mathcal{N}_1 < \mathcal{N}_2$ almost surely (a.s.). Hence,

$$\mathbb{E}_{\mathcal{N}_1}\left[\frac{\lambda(\mathcal{N}_1)}{\mathcal{N}_1}\middle|\, \mathcal{N}_1 \geq 1\right] < \mathbb{E}_{\mathcal{N}_2}\left[\frac{\lambda(\mathcal{N}_2)}{\mathcal{N}_2}\middle|\, \mathcal{N}_2 \geq 1\right],$$

i.e. the expected utilization $\hat{\rho}$ is strictly increasing in agents' participation probability p. □

Next, we analyze the $M/M/\mathcal{N}$ model with server priorities focusing on the FRFtS priority routing policy that allocates an incoming service request to the highest ranked, non-busy participating agent (if such agent does not exist, arriving customers wait in a single queue (pooled system) that is formed). In case there are priority *classes* with more than one agent in each of them, the firm routes an incoming service request to the highest priority class that has some non-busy participating agents, by picking one of them uniformly at random[1]. Similarly, if all participating agents are busy, the customers wait in the queue to be served.

To build intuition, assume that (in Step 1) the firm decided to split her total population of N agents into *two* priority classes with N_1 primary and N_2 secondary priority agents (the lower the index, the higher the priority class) such that $N_1 + N_2 = N$. Fix an arbitrary $n \in \{0, 1, \ldots, N\}$ and suppose that at the beginning of the period, $\{\mathcal{N}_j = n_j\}$ agents of priority class j participate and do not exit the system for the duration of the period, where $\mathcal{N}_j$ has support $\{0, \ldots, N_j\}$ for $j = 1, 2$, and such that $n_1 + n_2 = n$. Denote by $\mathcal{C}_1(t)$ and $\mathcal{C}_2(t)$ the number of customers occupying the primary and secondary agents respectively at time t such that $\mathcal{C}(t) = \mathcal{C}_1(t) + \mathcal{C}_2(t)$ is the total number of *customers* in the system. By noting that $\lambda(n) < n$, we can define

$$\pi(i, j; n) := \begin{cases} \lim_{t\to\infty} \mathbb{P}\left[\mathcal{C}_1(t) = i, \mathcal{C}_2(t) = j\right], & 0 \le i \le n_1,\ 0 \le j \le n_2 \\ \lim_{t\to\infty} \mathbb{P}\left[\mathcal{C}(t) = i + j\right], & i + j \ge n \end{cases} \tag{E.4}$$

as the joint steady state probability distribution function of this priority system (see Figure E.1 for an illustration of the underlying Markov Chain). We note that the stochastic processes $\{\mathcal{C}_1(t)\}_{t\ge 0}$, $\{\mathcal{C}_2(t)\}_{t\ge 0}$ are dependent, as the low priority servers will only be used once all the top priority servers are occupied. To the best of our knowledge, neither the marginal steady state distributions of $\{\mathcal{C}_1(t)\}_{t\ge 0}$ and $\{\mathcal{C}_2(t)\}_{t\ge 0}$ nor their joint steady state distribution are known. However, the steady state evolution of the (total) number of customers into the system $\{\mathcal{C}_1(t) + \mathcal{C}_2(t)\}_{t\ge 0}$ is given by the classical formulas of the $M/M/n$ model.

In order to find the expected utilizations $\hat{\rho}_1$ and $\hat{\rho}_2$ of the agents ranked in the primary and secondary priority class respectively, we proceed similarly to the case without priorities. In particular, we first characterize the steady state distribution $\pi(i, j; n)$ for every realizations n_1, n_2 with $n_1 + n_2 = n$, and then use it to calculate the expected long-run fraction of the time that a participating agent in each class is busy. In case there are more than two classes, we can compute the expected utilization of each priority class by an appropriate two priority classes re-partitioning. The following result provides a closed-form expression for the expected utilizations of each priority class for any number of priority classes formed.

Theorem 17 (Expected utilization of each priority class)**.** *(a) Assume that the firm forms two priority classes with $N_1 \in \{1, 2, \ldots, N-1\}$ primary priority agents and $N_2 = N - N_1$ secondary priority agents. Conditional that n_1, n_2 agents participate in each class with $n =$*

[1]This is indeed the approach currently used in practice by work-from-home contact centers, but we note that this is not entirely without loss of generality. It is outside the scope of this dissertation to investigate state dependent routing policies that could possibly allocate demand with positive probability to already busy participating agents depending on the queue length in front of them, and on the overall state of the system. We leave such cases to future research.

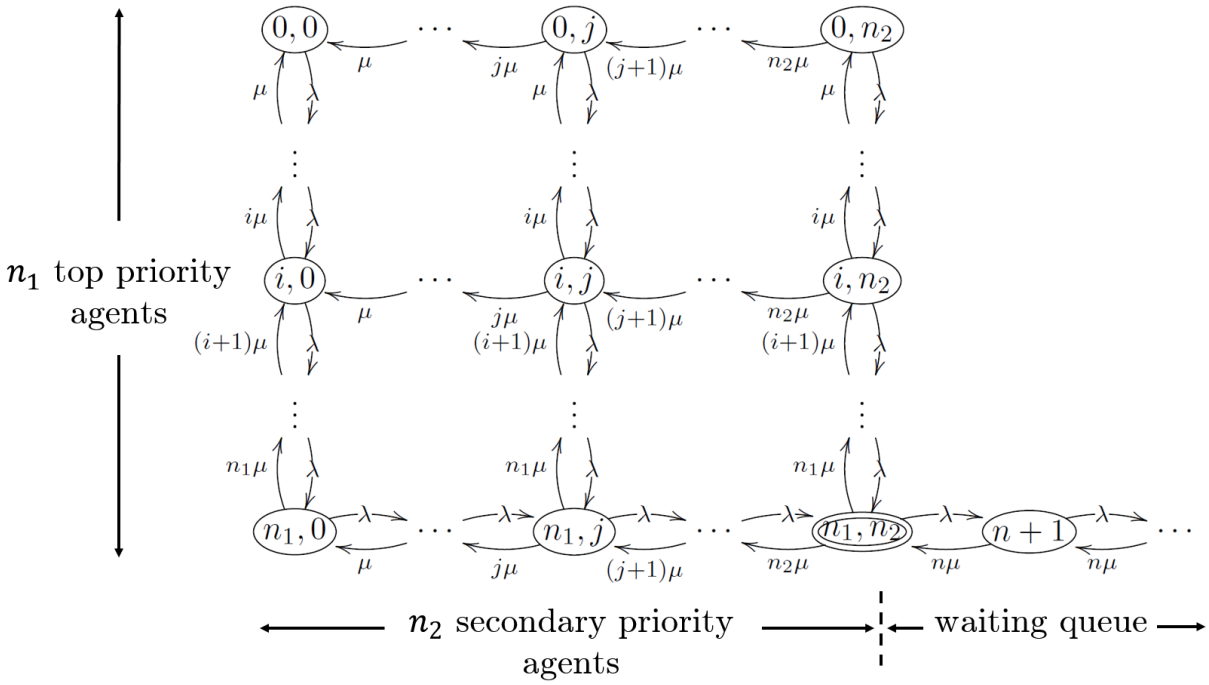


Figure E.1 Markov Chain of a stable $M/M/n$ model with ranked servers in two priority classes given that n_1 primary and $n_2 = n - n_1$ secondary priority servers have entered the system and serve demand at a fixed service rate $\mu > 0$.

$n_1 + n_2 \in \{1, 2, \ldots, N\}$ *and congestion-sensitive demand rate* $\lambda = \lambda(n)$ *with* $\lambda(n) < n$ *let*

$$\rho_1(n_1, n_2; \lambda) := \frac{\lambda}{n_1}\left(1 - B(n_1, \lambda)\right)\sum_{i=0}^{n}\hat{\pi}(i; n) + \sum_{k=n+1}^{\infty}\hat{\pi}(i; n) \tag{E.5}$$

$$\rho_2(n_1, n_2; \lambda) := \frac{\lambda - n_1 \cdot \rho_1(n_1, n_2; \lambda)}{n_2} \tag{E.6}$$

where $\hat{\pi}$ *is the steady state distribution* (E.1) *and* $B(n_1, \lambda)$ *is the Erlang-B loss formula of an* $M/M/n_1/n_1$ *system with mean traffic* λ*. Then, the expected utilization of the agents in the primary and secondary priority classes is given by*

$$\hat{\rho}_1 = \mathbb{E}_{\boldsymbol{\mathcal{N}}}\left[\rho_1(\mathcal{N}_1, \mathcal{N}_2; \lambda(\mathcal{N}_1 + \mathcal{N}_2))\right] \tag{E.7}$$

$$\hat{\rho}_2 = \mathbb{E}_{\boldsymbol{\mathcal{N}}}\left[\rho_2(\mathcal{N}_1, \mathcal{N}_2; \lambda(\mathcal{N}_1 + \mathcal{N}_2))\right] \tag{E.8}$$

(b) Consider a partition $\boldsymbol{N} = (N_1, \ldots, N_k)$ *with* $k > 2$ *priority classes and let* $\boldsymbol{\mathcal{N}} = (\mathcal{N}_1, \ldots, \mathcal{N}_k)$ *be the ex ante vector of the number of participating agents in each class. Then, the expected utilization of class* j *out of* k *is given by* $\hat{\rho}_{j:k} = \mathbb{E}_{\boldsymbol{\mathcal{N}}}\left[\frac{\lambda_{j:k}(\boldsymbol{\mathcal{N}}; \lambda)}{\mathcal{N}_j}\middle|\, \mathcal{N}_j \geq 1\right]$*, where*

$$\lambda_{j:k}(\boldsymbol{\mathcal{N}}; \lambda) := \begin{cases} \mathcal{N}_1 \cdot \rho_1\left(\mathcal{N}_1, \sum_{i=2}^{k}\mathcal{N}_i; \lambda\right), & j = 1 \\ \left(\sum_{i=1}^{j}\mathcal{N}_i\right) \cdot \rho_1\left(\sum_{i=1}^{j}\mathcal{N}_i, \sum_{i=j+1}^{k}\mathcal{N}_i; \lambda\right) - \left(\sum_{i=1}^{j-1}\mathcal{N}_i\right) \cdot \rho_1\left(\sum_{i=1}^{j-1}\mathcal{N}_i, \sum_{i=j}^{k}\mathcal{N}_i; \lambda\right), & j = 2, \ldots, k-1 \\ \lambda - \sum_{i=1}^{k-1}\mathcal{N}_i \cdot \rho_1\left(\sum_{i=1}^{k-1}\mathcal{N}_i, \mathcal{N}_k; \lambda\right), & j = k \end{cases}$$

Proof of Theorem 17. (a) Assume that the firm forms two priority classes with N_1 primary priority agents where $N_1 \in \{1, 2, \ldots, N-1\}$ and $N_2 = N - N_1$ secondary priority agents. Conditional that n_1, n_2 agents participate in each class with $n = n_1 + n_2 \in \{1, 2, \ldots, N\}$ and

congestion-sensitive demand rate $\lambda = \lambda(n)$ with $\lambda(n) < n$ we have the following. (Note that if one class has no participating agents then its (realized) utilization is zero by definition, and the other class has n agents so its (realized) utilization is $\frac{\lambda}{n}$ from the traditional $M/M/n$ model.) By definition the utilization of the primary and secondary priority classes are the long-run average fraction of time an agent of each class is occupied respectively:

$$\rho_1(n_1, n_2; \lambda) := \sum_{j=0}^{n_2} \sum_{i=1}^{n_1-1} \frac{i}{n_1} \pi(i, j; n_1 + n_2) + \sum_{k=n}^{\infty} \hat{\pi}(k; n) \tag{E.9}$$

$$\rho_2(n_1, n_2; \lambda) := \sum_{i=0}^{n_1} \sum_{j=1}^{n_2-1} \frac{j}{n_2} \pi(i, j; n_1 + n_2) + \sum_{k=n}^{\infty} \hat{\pi}(k; n) \tag{E.10}$$

where π and $\hat{\pi}$ are the steady state distributions (E.4) and (E.1). Observe that (E.3) in the proof of Lemma 11 implies that $\rho_1(n_1, 0; \lambda) = \frac{\lambda}{n_1}$ and $\rho_1(0, n_2; \lambda) = \frac{\lambda}{n_2}$.

Reading Figure E.1 row by row, the following flow-balance equations for the stationary distribution π can be derived when $n_1, n_2 \geq 1$:

$$\begin{aligned}
(\lambda + j)\,\pi(0, j; n) &= \pi(1, j; n) + (j+1)\,\pi(0, j+1; n), && 0 \leq j \leq n_2 - 1 \\
(\lambda + n_1)\,\pi(0, n_2; n) &= \pi(1, n_2; n) \\
(\lambda + i + j)\,\pi(i, j; n) &= \lambda\pi(i-1, j; n) + (i+1)\,\pi(i+1, j; n) + (j+1)\,\pi(i, j+1; n), && 1 \leq i \leq n_1 - 1, \\
& && 0 \leq j \leq n_2 - 1 \\
(\lambda + i + n_2)\,\pi(i, n_2; n) &= \lambda\pi(i-1, n_2; n) + (i+1)\,\pi(i+1, n_2; n), && 1 \leq i \leq n_1 - 1 \\
(\lambda + n_1)\,\pi(n_1, 0; n) &= \lambda\pi(n_1 - 1, 0; n) + \pi(n_1, 1; n) \\
(\lambda + n_1 + j)\,\pi(n_1, j; n) &= \lambda\left[\pi(n_1 - 1, j; n) + \pi(n_1, j-1; n)\right] + (j+1)\,\pi(n_1, j+1; n), && 1 \leq j \leq n_2 - 1 \\
(\lambda + n)\,\pi(n_1, n_2; n) &= \lambda\left[\pi(n_1 - 1, n_2; n) + \pi(n_1, n_2 - 1; n)\right] + n\,\hat{\pi}(n+1; n) \\
\pi(0, 0; n) \equiv \hat{\pi}(0; n) &= \left(\sum_{r=0}^{n-1} \frac{\lambda^r}{r!} + \frac{\lambda^n}{n!\left(1 - \frac{\lambda}{n}\right)}\right)^{-1} \\
\pi(i, j; n) \equiv \hat{\pi}(i + j; n) &= \frac{n^n \left(\frac{\lambda}{n}\right)^{i+j}}{n!}\,\hat{\pi}(0; n) && n_1 + n_2 \leq i + j < \infty \\
\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi(i, j; n) &= 1
\end{aligned} \tag{E.11}$$

Define the function $p(i) := \sum_{j=0}^{n_2} \pi(i, j; n_1 + n_2)$ for each $i = 0, 1, \ldots, n_1$ and observe that $\rho_1(n_1, n_2; \lambda) = \sum_{i=1}^{n_1-1} \frac{i}{n_1} p(i) + \sum_{k=n}^{\infty} \hat{\pi}(k; n)$. Then, $p(i)$ can be viewed as the steady state distribution of the Markov Chain illustrated in Figure E.2.

The flow-balance equations are

$$\left.\begin{aligned}
\lambda \cdot p(0) &= p(1) \\
(\lambda + i) \cdot p(i) &= \lambda \cdot p(i-1) + (i+1) \cdot p(i+1), \qquad i = 1, \ldots, n_1 - 1
\end{aligned}\right\}$$

and imply that $p(i) = \frac{\lambda^i}{i!} \cdot p(0)$, $i = 1, \ldots, n_1$. The key to find $p(0)$ without referring back to (E.11) is to observe that the total system operates as the $M/M/n$, hence $\sum_{i=0}^{n_1} p(i) =$

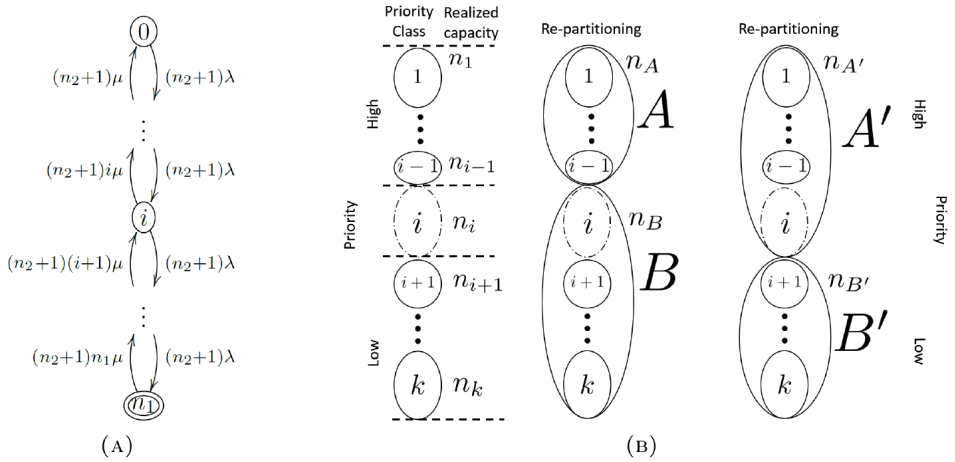


Figure E.2 (A) Markov Chain for the evolution of $p(i)$ (Theorem 17a). (B) Reduction to two priority classes by re-partitioning when there $k \geq 3$ priority classes.

$\sum_{i=0}^{n_1+n_2} \hat{\pi}(i;\, n)$. That is, $p(0) + \sum_{i=1}^{n_1} \frac{\lambda^i}{i!} \cdot p(0) = \sum_{i=0}^{n_1+n_2} \hat{\pi}(i;\, n)$, or

$$p(0) = \frac{\sum_{i=0}^{n_1+n_2} \hat{\pi}(i;\, n)}{\sum_{i=0}^{n_1} \frac{\lambda^i}{i!}}$$

Then, by the definition of the Erlang-B formula $B(n_1,\, \lambda) := \frac{\frac{\lambda^{n_1}}{n_1!}}{\sum_{i=0}^{n_1} \frac{\lambda^i}{i!}}$ we get

$$\begin{aligned}
\rho_1(n_1,\, n_2;\, \lambda) &= \sum_{i=1}^{n_1} \frac{i}{n_1} p(i) + \sum_{k=n+1}^{\infty} \hat{\pi}(k;\, n) \\
&= \sum_{i=1}^{n_1} \frac{i}{n_1} \cdot \frac{\lambda^i}{i!} \cdot \left\{ \frac{\sum_{i=0}^{n_1+n_2} \hat{\pi}(i;\, n)}{\sum_{i=0}^{n_1} \frac{\lambda^i}{i!}} \right\} + \sum_{k=n+1}^{\infty} \hat{\pi}(k;\, n) \\
&= \frac{\lambda}{n_1} \cdot (1 - B(n_1,\, \lambda)) \sum_{i=0}^{n_1+n_2} \hat{\pi}(i;\, n) + \sum_{k=n+1}^{\infty} \hat{\pi}(k;\, n)
\end{aligned}$$

Note that $\hat{\pi}$ is known from the $M/M/n$ model. Knowing $\rho_1(n_1,\, n_2;\, \lambda)$ we can find $\rho_2(n_1,\, n_2;\, \lambda)$ by the conservation of the effective traffic rate in each priority class as follows. Let $\lambda_1 = n_1 \cdot \rho_1(n_1,\, n_2;\, \lambda)$ and $\lambda_2 = n_2 \cdot \rho_2(n_1,\, n_2;\, \lambda)$ be the effective traffic rates in the primary and secondary priority classes respectively. Then, $\lambda_1 + \lambda_2 = \lambda$, or $n_1 \cdot \rho_1(n_1,\, n_2;\, \lambda) + n_2 \cdot \rho_2(n_1,\, n_2;\, \lambda) = \lambda$, which implies

$$\rho_2(n_1,\, n_2;\, \lambda) = \frac{\lambda - n_1 \cdot \rho_1(n_1,\, n_2;\, \lambda)}{n_2}, \qquad n_2 \geq 1$$

Then, for a congestion-sensitive demand with customer expected utility (2.1), the expected utilization of the agents in the primary and secondary priority classes are given by

$$\hat{\rho}_1(N_1,\, N_2;\, \lambda) = \mathbb{E}_{\mathcal{N}}\left[\rho_1(\mathcal{N}_1,\, \mathcal{N}_2;\, \lambda(\mathcal{N}_1 + \mathcal{N}_2))\right]$$

$$\hat{\rho}_2(N_1,\, N_2;\, \lambda) = \mathbb{E}_{\mathcal{N}}\left[\rho_2(\mathcal{N}_1,\, \mathcal{N}_2;\, \lambda(\mathcal{N}_1 + \mathcal{N}_2))\right]$$

(b) Suppose that there are $k \geq 3$ priority classes and fix a priority class i, for $i = 2,\, 3, \ldots,\, k-1$.

Observe that traffic is sent to class i if and only if the agents of higher priority are all busy (see also Figure E.2B). We show the statement by conditioning on the realized number of participating agents in each priority class to find the realized values of utilizations for each class. Condition that $n_1, \ldots, n_k$ agents have participated in each class and let $\sum_{i=1}^{k} n_i = n$. The congestion-sensitive, arriving traffic into the system is then $\lambda = \lambda(n)$. We re-partition the initial system into class A (primary) with the top $n_A := \sum_{j=1}^{i-1} n_j$ agents and class B (secondary) with the remaining $n_B := \sum_{j=i}^{k} n_j$ agents (see also Figure E.2B). Applying part (a) we have that the utilization of class A into this new two priority classes system is $\rho_1(n_A, n_B; \lambda)$ and the effective traffic rate into class A is $\lambda_A = n_A \cdot \rho_1(n_A, n_B; \lambda)$. Re-partition again the initial system into a new class A' (primary) with the top $n_{A'} := n_A + n_j$ agents and a new class B' (secondary) with the remaining $n_{B'} := \sum_{j=i+1}^{k} n_j$ agents (see also Figure E.2B). Applying part (a) we have that the utilization of class A' into this new two priority classes system is $\rho_1(n_{A'}, n_{B'}; \lambda)$ and the effective traffic rate into class A is $\lambda_{A'} = n_{A'} \cdot \rho_1(n_{A'}, n_{B'}; \lambda)$. We then have that the effective arrival rate into the initial class j is equal to $\lambda_{j:k} := \lambda_{A'} - \lambda_A$. We define $\lambda_{1:k} := n_1 \cdot \rho_1\left(n_1, \sum_{j=2}^{k} n_j; \lambda\right)$ and $\lambda_{k:k} := \lambda - \sum_{j=1}^{k-1} n_j \cdot \rho_1\left(\sum_{j=1}^{k-1} n_j, n_k; \lambda\right)$, so that $\lambda_{1:k} = \lambda$.

Then, the utilization of class j is equal to $\rho_{j:k} = \frac{\lambda_{j:k}}{n_j}$ where

$$\lambda_{j:k} := \begin{cases} n_1 \cdot \rho_1\left(n_1, \sum_{j=2}^{k} n_j; \lambda\right), & i = 1 \\ \left(\sum_{j=1}^{i} n_j\right) \cdot \rho_1\left(\sum_{j=1}^{i} n_j, \sum_{j=i+1}^{k} n_j; \lambda\right) - \left(\sum_{j=1}^{i-1} n_j\right) \cdot \rho_1\left(\sum_{j=1}^{i-1} n_j, \sum_{j=i}^{k} n_j; \lambda\right), & i = 2, \ldots, k-1 \\ \lambda - \sum_{j=1}^{k-1} n_j \cdot \rho_1\left(\sum_{j=1}^{k-1} n_j, n_k; \lambda\right), & i = k \end{cases}$$

The statement follows by taking the expectation over the vector $\boldsymbol{\mathcal{N}} = (\mathcal{N}_1, \ldots, \mathcal{N}_k)$ of the number of participating agents in each priority class. □

The simplified expressions provided in Theorem 17(a) can be used to find the utilizations of each priority class of an $M/M/n$ system with n servers ranked in k priority classes, for each $k \in \{2, \ldots, n\}$. We show that the expression for the utilization of the primary priority class (E.5) for $n_1 := n$ reduces to the known utilization expression $\frac{\lambda}{n}$ of the $M/M/n$ system without priorities (i.e. $k := 1$). For an intuition for $k \geq 3$ classes, suppose that the firm has split its population into a 3-partition with N_1, N_2 and $N_3 = N - N_1 - N_2$ agents respectively. Assume further that n_A, n_B, $n_C := n - n_A - n_B$ agents have participated in each of them respectively. To find the utilization of the top priority class A composed of n_A agents, we split the system into two new priority classes: the top n_A (primary) and the rest $n_B + n_C$ (secondary) agents. Using Theorem 17(a) we calculate the utilization of the class A and its effective arrival rate λ_A. Next, we re-partition the system into the top $n_A + n_B$ (primary) and the rest n_C (secondary) agents. Using Theorem 17(a) we can find the utilization of the new top priority class and its effective arrival rate λ'_A. By the conservation of traffic, the effective arrival rate into class B is the difference $\lambda'_A - \lambda_A$. A similar procedure outlined in Theorem 17(b) gives the utilization and effective arrival rate of class C as well.

Theorem 17 makes no assumption on the distribution of the number of participating agents. In the special case where only a fraction of the top ranked agents participate according to a Binomial distribution, an analog of Lemma 11(b) holds for *every* priority class. Intuitively, as more agents are expected to participate, the incoming traffic to each priority class and the

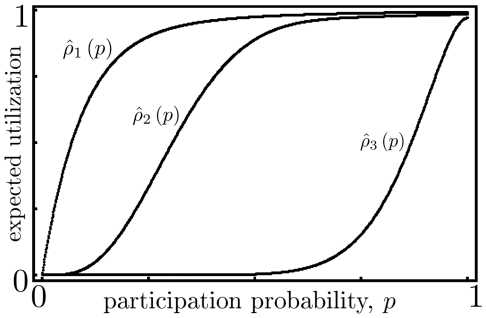


Figure E.3 Consider a partition of $N = 12$ agents into three priority classes with capacity of $N_1 = 3$, $N_2 = 7$ and $N_3 = 2$ agents, respectively. Assume that $V = 10$, $c = 2$ and endogenous demand λ that solves (2.1). All else equal, the expected utilizations of all priority classes increase in agents' participation probability, when only a fraction of the top ranked agents participate according to a Binomial distribution.

number of agents of each class stochastically increase at a rate such that the utilization of each priority class stochastically increases as well. Hence, the expected utilization of each priority class increases in agents' participation probability. Figure E.3 summarizes this insight. Indeed, as we show in §2.4 such a Binomial distribution arises in equilibrium and characterizes the voluntary participation strategy of the work-from-home agents.

E.3 Proofs

Proof of Lemma 3. We focus on symmetric pure strategies. If the firm is disregarding any priority classes (i.e. $k := 1$), it allocates equal amount of traffic to any agent who participates, irrespective of his ability rank-order. That is, the expected utility of an agent with ability a is $u(a; \tilde{p}) = \frac{w \cdot \lambda(N)}{N} < c_p$, due to assumption (2.4). Hence, no agent finds it rational to participate, i.e. agents' participation probability satisfies $p^* = 0$. □

Proof of Theorem 6. (a) and (b) Consider a population of N agents and suppose that the firm has decided on a relative ranking scheme of k ($k = 2, \ldots, N$) ranks each with N_j ($j = 1, \ldots, k$) agents so that $\sum_{j=1}^{k} N_j = N$. Incoming demand is then allocated according to FRFtS routing among the k priority classes formed, for any number of participating agents from each rank.

In order to make a rational participation decision, agent i ($i = 1, \ldots, N$) forms beliefs which are conjectures (in the form of a distribution) about the equilibrium participation actions of the others. We are interested in examining whether a symmetric pure equilibrium exists under symmetric and consistent beliefs, while excluding correlated beliefs. Suppose that based on his beliefs agent i conjectures that each other agent $z \neq i$ participates with probability $\tilde{p}_z$. Focusing on symmetric equilibria and since agent i does not know the ability of the others, we have $\tilde{p}_z = \tilde{p}$ for all $z \neq i$, i.e. $\tilde{p}_z$ does not depend on agent z's ability (type) a_z.

Taking expectation over the realized number of participating agents of any rank, the expected utilization of a participating agent ranked in the jth priority class is

$$\hat{\rho}_j(\tilde{p}) = \sum_{0 \leq n_1 \leq N_1} \cdots \sum_{0 \leq n_k \leq N_k} \binom{N}{\sum_{j=1}^{k} n_j} \cdot \tilde{p}^{\sum_{j=1}^{k} n_j} \cdot (1 - \tilde{p})^{N - \sum_{j=1}^{k} n_j} \cdot \frac{\lambda_{j:k}(\boldsymbol{n}; \lambda)}{n_j}, \quad \text{(E.12)}$$

where $\lambda_{j:k}(\boldsymbol{n};\lambda)$ is the effective arrival rate to priority class j given by Theorem 17(b), and $\boldsymbol{n}=(n_1,\ldots,n_k)$. Note that the above expression does not depend on the ability of any agent. Also, the incoming traffic rate λ in (E.12) is congestion-sensitive so it depends on the realized total number of participating agents $n=\sum_{j=1}^{k}n_j$ and solves (2.1) as described in §2.3.

Without loss of generality, we express the k priority classes resulting in k *distinct* expected utilizations as N expected utilizations with some of them to be identical. The expected utility of agent i who has ability a and conjectures that all others participate w.p. $\tilde{p}$ is

$$\begin{aligned} u(a;\tilde{p}) &= \sum_{j=1}^{N} w\,\hat{\rho}_j(\tilde{p})\cdot\mathbb{P}\left[a \text{ is ranked } j\text{th out of } N\right] \\ &= \sum_{j=1}^{N}\hat{\rho}_j(\tilde{p})\cdot\left\{F_{N-j:N-1}(a)-F_{N-j+1:N-1}(a)\right\} \\ &= \sum_{j=1}^{N} w\left(\hat{\rho}_j(\tilde{p})-\hat{\rho}_{j+1}(\tilde{p})\right)\cdot F_{N-j:N-1}(a), \end{aligned} \tag{E.13}$$

where we define $\hat{\rho}_{N+1}:=0$. Note that there are only k *distinct* values of expected utilizations $\hat{\rho}_j$ in (E.13); all agents ranked in the same priority class are equally utilized. We also have that $u(\cdot;\tilde{p})$ is continuous and strictly increasing, since $F(\cdot)$ is strictly increasing on $[0,1]$ (by assumption) which implies that the order-statistics distribution is also strictly increasing.

Let a_{min} be the solution (if it exists) to the equation $u(a;\tilde{p})=c_p$. If $u(a;\tilde{p})>c_p$ for all $a\in[0,1]$, we set $a_{min}:=0$; if $u(a;\tilde{p})<c_p$ for all $a\in[0,1]$, we set $a_{min}:=1$. The strict monotonicity of the expected utility implies that if an agent with ability a participates, all other agents with abilities $a'\geq a$ participate as well. Due to symmetry there is a unique (global) participation threshold $a_{min}=a_{min}(\hat{\boldsymbol{\rho}};\tilde{p})$ which is the same across all agents. In the following three steps we show that the agents have a *unique* self-confirming equilibrium belief defined as $\tilde{p}=1-F(a_{min})$.

Step 1: If agents participate according to a threshold strategy with a probability $\tilde{p}$, the expected utilization $\hat{\rho}_j(\tilde{p})$ of every priority class j strictly increases in $\tilde{p}$. The threshold participation strategy of the agents imply that the participating agents fill up a fraction of the top positions (without gaps). From Lemma 11 we know that the expected utilization of a system without priorities (or the expected number of participating agents under a threshold strategy) strictly increases in $\tilde{p}$. Hence, as $\tilde{p}$ increases, the expected utilization $\hat{\rho}_j(\tilde{p})$ of every priority class j strictly increase as well.

Step 2: The unique participation threshold $a_{min}=a_{min}(\hat{\boldsymbol{\rho}};\tilde{p})$ is strictly decreasing in the belief $\tilde{p}$. Indeed, an agent with ability a finds it rational to participate in the service contest w.p. $\tilde{p}$ iff his expected utility $u(a;\tilde{p})$ of participating w.p. $\tilde{p}$ and being placed at any of the available rankings when competing with $N-1$ other agents covers his participation cost c_p:

$$\begin{aligned} \sum_{j=1}^{N} w\,\hat{\rho}_j(\tilde{p})\cdot\left\{\binom{N-1}{N-j}\tilde{p}^{N-j}(1-\tilde{p})^{j-1}\right\} &= \sum_{j=1}^{N} w\,\hat{\rho}_j(1-\tilde{p})\cdot\left\{\binom{N-1}{j-1}\tilde{p}^{j-1}(1-\tilde{p})^{N-j}\right\} \\ &> c_p \end{aligned}$$

The equation follows by the pigeonhole principle and the terms in brackets are the PDF of

the Binomial $(N-1,\,\tilde{p})$ distribution. Since $\hat{\rho}_j\,(1-\tilde{p})$ strictly decreases in $\tilde{p}$ for all rankings $j=1,\ldots,N$, due to stochastic dominance we have that $u\,(a;\,\tilde{p})$ strictly decreases in $\tilde{p}$ (or $\frac{\partial u}{\partial \tilde{p}}\,(a_{min},\,\tilde{p})<0$). Further, due to the strict monotonicity of agents' expected utility in ability, we have that $\frac{\partial u}{\partial a_{min}}\,(a_{min},\,\tilde{p})>0$. Assuming differentiability the Envelope Theorem

$$\frac{du}{d\tilde{p}}\,(a,\tilde{p})\bigg|_{a=a_{min}}=\underbrace{\frac{\partial u}{\partial a_{min}}\,(a_{min}\,(\tilde{p})\,,\tilde{p})}_{>0}\cdot\frac{\partial a_{min}}{\partial \tilde{p}}\,(\tilde{p})+\underbrace{\frac{\partial u}{\partial \tilde{p}}\,(a_{min}\,(\tilde{p})\,,\,\tilde{p})}_{<0}=0$$

implies that $\frac{\partial a_{min}}{\partial \tilde{p}}\,(\tilde{p})$ must be positive.

Step 3: There exists a unique solution $\tilde{p}^*\in[0,1]$ to the equation $\tilde{p}=1-F\,(a_{min}\,(\tilde{p}))$. Indeed, since F is assumed continuous and strictly increasing in ability, the function $k\,(\tilde{p}):=1-F\,(a_{min}\,(\tilde{p}))-\tilde{p}$ is continuous and strictly decreasing in $\tilde{p}\in[0,1]$. If $\tilde{p}=0$, then every agent who enters will cover his participation cost with positive probability and hence $a_{min}\,(0)>0$ and $k\,(0)=1-F\,(a_{min}\,(0))>0$. If $\tilde{p}=1$, then this can not be a SCE since in that case the condition (2.4) implies that every agent who enters does not cover his participation cost. So, the decision to participate is not rational. Hence, if $\tilde{p}=1$, we have $a_{min}\,(1)<1$ and $k\,(1)=-F\,(a_{min}\,(1))<0$. By the Intermediate Value Theorem, a unique $\tilde{p}^*$ must exist such that $k\,(\tilde{p}^*)=0$. □

Proof of Theorem 7. Suppose that the agents form an *ex ante* belief $\tilde{p}$ on the *ex post* participation probability p^* of the others. At the unique SCE shown in Theorem 6 we require $\tilde{p}=p^*$. Hence, the number of participating agents in equilibrium: $\mathcal{N}^*=\sum_{i=1}^{N}\mathbb{1}_{\{\text{agent } i \text{ participates}\}}$ is a sum of independent Bernoulli trials with success probability p^*. By definition, we have that $\mathcal{N}^*$ follows the Binomial distribution with parameters $(N,\,p^*)$.

Suppose that the equilibrium number of participating agents $\mathcal{N}^*\sim$ Binomial $(N,\,p_N)$, where $p_N:=1-F\left(a_{min}^N\right)$ denotes the participation probability chosen by the agents in equilibrium given a population of size N. Define $\overline{m}:=\max\left\{n\in\mathbb{N}:\ \frac{w\Lambda}{n}\geq c_p\right\}=\max\left\{1,\,\left\lceil\frac{w\Lambda}{c_p}\right\rceil-1\right\}$, and note that $\overline{m}$ is an exogenously fixed constant and does not depend on agents' population size N. By its definition $\overline{m}$ represents an endogenously arising cap on the number of top ranked agents that the firm can route a non-zero fraction of the available demand (if more than $\overline{m}$ agents participate they should not be utilized). That is, $\overline{m}$ represents a form of a "toll" that the firm should set to account for the strategic behavior of her agents when the population size is large (see also Gurvich et al. (2015) who show in a similar self-scheduling setting that imposing a cap on the agents who can are allowed to enter can be optimal). Then, for each $k\geq\overline{m}$ we have that no agent participates since their expected utility $w\lambda\,(N)-kNc_p<w\Lambda N-kNc_p<0$. Hence, the *expected* number of participating agents in SCE $N\cdot p_N$ can not exceed $\overline{m}$. That is, $p_N=O\left(\frac{1}{N}\right)$ and $\lim_{N\to\infty}p_N=0$ (or equivalently $\lim_{N\to\infty}a_{min}^N=1$).

(*Existence*) Since $p_N=O\left(\frac{1}{N}\right)$, there exists a constant $n_\infty>0$ which is independent of N such that $\lim_{N\to\infty}N\cdot p_N=n_\infty$. Hence, the Poisson Limit Theorem can be applied and implies the convergence of the Binomial $(N,\,p_N)$ to a Poisson distribution with parameter n_∞. We next show how to determine the value of the limit n_∞.

Another application of the Poisson Limit Theorem implies that

$$\lim_{N\to\infty}\binom{N-1}{j-1}p_N^{j-1}\,(1-p_N)^{N-j}=e^{-n_\infty}\frac{(n_\infty)^{j-1}}{(j-1)!}$$

Observe that scaling the nominal rate to $\Lambda\cdot N$ and agents' service rate to N stochastically decreases the realized utilizations of each priority class. Then, an asymptotic version of the (IR) constraint (2.5) gives that n_∞ is the solution of

$$\sum_{j=1}^{\overline{m}} w\hat{\rho}_j^{\text{Poisson}}\left(n_\infty\right)\cdot e^{-n_\infty}\frac{(n_\infty)^{j-1}}{(j-1)!}=c_p, \tag{E.14}$$

where $\hat{\rho}_j^{\text{Poisson}}\left(n_\infty\right)$ denotes here the expected utilization of a class j over the number of participating agents who follow Poisson $\left(n_\infty\right)$, for $j\geq 1$.

(*Uniqueness*) For each $z\in[0,\,\overline{m}]$, $h\left(z\right)=\sum_{j=0}^{\infty}w\hat{\rho}_j^{\text{Poisson}}\left(z\right)e^{-z}\frac{z^{j-1}}{(j-1)!}$ is a continuous and strictly decreasing function of z defined on the interval $[0,\,\overline{m}]$. This implies that $h\left(\cdot\right)$ has a continuous inverse. □

Lemma 12. *For each $1\leq j<k\leq N$ the difference $\frac{F_{N-j:N-1}(a)}{j}-\frac{F_{N-k:N-1}(a)}{k}$ is single crossing zero from positive to negative. Hence, there is a unique $j^*=j^*\left(a,\,N\right)$ such that $j^*=\arg\max_{1\leq j\leq k}\left\{\frac{F_{N-j:N-1}(a)}{j}\right\}$ for each $a\in[0,\,1]$.*

Proof of Lemma 12. Set $j:=N-s$, $s=0,\ldots,N-1$ which counts the ranking positions in reverse order from the lowest to the highest one. Due to stochastic dominance of order statistics it suffices to show the statement for two consecutive positions: $\Delta=\Delta\left(a\right)=\frac{F_{j:N-1}(a)}{N-j}-\frac{F_{j+1:N-1}(a)}{N-j-1}$, $\forall a\in[0,\,1]$. Using the integral representation (B.1) we have the expression:

$$\Delta\left(a\right)=\frac{N}{\left(N-j\right)\left(N-j-1\right)}\left(\begin{array}{c}N-2\\ j\end{array}\right)\int_0^{F(a)}x^{j-1}\left(1-x\right)^{N-2-j}\left\{j-Nx\right\}dx$$

Note that the function $\widehat{G}\left(x\right):=j-Nx$ is continuous and $\varphi\left(x\right):=x^{j-1}\left(1-x\right)^{N-2-j}$ is an integrable function that does not change sign on the interval $(0,\,F\left(a\right))$. From the First Mean Value Theorem for integration there exists $\check{x}$ in $[0,\,F\left(a\right)]$ such that $\int_0^{F(a)}x^{j-1}\left(1-x\right)^{N-2-j}\left\{j-Nx\right\}dx=\left(j-N\,\check{x}\right)\cdot\int_0^{F(a)}x^{j-1}\left(1-x\right)^{N-2-j}dx$. Now, the sign $\left[\Delta\right]$ is determined by the sign $\left[j-N\,\check{x}\right]$. Hence, there is a unique $j^*=j^*\left(a,\,N\right)$ such that sign $\left[\Delta\right]$ is negative for $j<j^*$, zero at j^* and positive for $j>j^*$. For any k, $k\in\{2,\ldots,N-1\}$ the latter implies that $\frac{F_{N-k-1:N-1}(a)}{k+1}-\frac{F_{N-k:N-1}(a)}{k}$ is first increasing and then decreasing. □

Proof of Theorem 8. The unique threshold participation strategy of the agents implies that a number of agents ranked in the top positions would participate in equilibrium. We also have that the expected utilization of each priority class strictly increases in agents' equilibrium participation probability. Further, agents' participation threshold is global, i.e. it does not depend on the ability of each agent. Hence, firm's problem (2.7) is equivalent to finding the priority classes partition (in terms of number of *distinct* expected utilization values) that maximizes agents' equilibrium participation probability p (or equivalently finding the partition that minimizes agents' equilibrium participation threshold $a_{min}=a_{min}\left(\hat{\boldsymbol{\rho}}\right)$), subject to the rest of

the constraints.

We define the *normalized expected utilization differentials*

$$\hat{\delta}_j := j \cdot (\hat{\rho}_j - \hat{\rho}_{j+1}), \qquad j = 1, \ldots, N$$

and express firm's problem using such differences in expected utilizations between classes. Observe that the monotonicity constraints in (2.7) take the form of the non-negativity constraints $\hat{\delta}_j \geq 0$, $j = 1, \ldots, N$. Taking expectation wrt the number of participating agents $\mathcal{N}$ the budget constraint (2.7) can be written as $\sum_{j=1}^{N} \hat{\delta}_j = \lambda$. Then, the deltas chosen by the firm determine agents' participation threshold through the equation

$$\sum_{j=1}^{N} w\hat{\delta}_j \left\{ \frac{F_{N-j:N-1}(a_{min})}{j} \right\} = c_p \tag{E.15}$$

The terms in brackets are the "weights", which are unimodal wrt j for all $a_{min} \in [0, 1]$ as shown in Lemma 12. Maximizing the LHS of (E.15) gives the highest participation cost that can be supported by a given demand allocation $\hat{\boldsymbol{\delta}}$. That is, to find the minimum participation threshold over any demand allocation $\hat{\boldsymbol{\delta}}$, one has to set the expected utilization difference $\hat{\delta}_{j^*}$ equal to its maximum value for the index j^* with the largest coefficient multiplying $\hat{\delta}_{j^*}$ in (E.15), making all the rest $\hat{\delta}_j$'s equal to zero. That is, it is optimal to form two priority classes with

$$N_1^* := \arg \max_{1 \leq j \leq N} \left\{ \frac{F_{N-j:N-1}(a_{min})}{j} \right\}$$

top priority agents equally utilized and offer a strictly lower utilization to $N_2^* := N - N_1^*$ lower priority agents according to FRFtS routing (see Theorem 17). FRFtS routing implies that the firm is routing demand only to participating agents so that all participating agents receive a positive utilization, and any non-participating agents are not utilized. Hence, we have

$$\hat{\rho}_1 \cdot F_{N-N_1^*:N-1}(a_{min}) + \hat{\rho}_2 \cdot \left\{ 1 - F_{N-N_1^*:N-1}(a_{min}) \right\} = \frac{c_p}{w}$$

The value of a_{min} that solves the above equation determines the maximum participation probability $p^* = 1 - F(a_{min})$ that can be supported for a given incoming demand rate and the rest exogenous parameters of the system. That is, for available demand λ and participation/wage ratio $\frac{c_p}{w}$, the firm can at most induce a number of participating agents distributed as $\mathcal{N} \sim \text{Binomial}(N, p^*)$. □

Proof of Theorem 9. (a) Theorem 6 implies that the ability threshold a_{min} is increasing in agents' participation cost c_p, i.e. for each $\hat{c}_p < c_p$ we will have that $\hat{a}_{min}(\hat{c}_p) < a_{min}(c_p)$. To show that $N_1^* := \arg\max_{1 \leq j \leq N} \left\{ \frac{F_{N-j:N-1}(a_{min})}{j} \right\}$ is (weakly) decreasing in c_p, it suffices to show that N_1^* is (weakly) decreasing in a_{min}. Lemma 12 shows that for a given a_{min}, the weights $A_j(a_{min}) = \frac{F_{N-j:N}(a_{min})}{j}$ have a unique maximum wrt j. Further, the function $\widetilde{f}(a_{min}) := \max_{1 \leq j \leq N} A_j(a_{min})$ is increasing in a_{min}. Hence, $\arg\max_{1 \leq j \leq N} A_j(\hat{a}_{min}) \geq \arg\max_{1 \leq j \leq N} A_j(a_{min})$, which implies that $N_1^*(\hat{a}_{min}) \geq N_1^*(a_{min})$.

(b) We first prove that $N_1^* = N_1^*(N)$ is weakly increasing in N. We have shown that for a fixed a_{min}, $N_1^*(c_p, a_{min})$ is (weakly) decreasing in c_p. We further have that all else equal, c_p

increases in a_{min}. It suffices to show that a_{min} is weakly increasing in agents' population N. Indeed, the stochastic dominance of the order statistics distribution $F_{i+1:\,N+1}(x) \leq F_{i:\,N}(x)$ for all N, $i \in \{1,\ldots,N-1\}$, and x in its domain imply that for a fixed N_1^* and c_p, and for each $\hat{N} < N$ we will have that $\hat{a}_{min}\left(\hat{N}\right) \leq a_{min}(N)$. Hence, $N_1^* = N_1^*(N)$ is weakly increasing in N.

(c) Arguing similarly to Theorem 8 and with the scaling of Theorem 7, as $N \to \infty$ the asymptotic (IR) constraint is given by (E.14) and the optimal partition contains two priority classes with N_1^* top priority agents given by the unique solution to the system:

$$\left.\begin{array}{c} N_1^* = \arg\max_{1\leq j\leq \overline{m}} \left\{ \frac{1}{j} \sum_{k=1}^{j} e^{-n_\infty} \frac{(n_\infty)^{k-1}}{(k-1)!} \right\} \\ \hat{\rho}_1^{\text{Poisson}} \cdot \left(\sum_{k=1}^{N_1^*} e^{-n_\infty} \frac{(n_\infty)^{k-1}}{(k-1)!} \right) + \hat{\rho}_2^{\text{Poisson}} \cdot \left(\sum_{k=N_1^*+1}^{\overline{m}} e^{-n_\infty} \frac{(n_\infty)^{k-1}}{(k-1)!} \right) = \frac{c_p}{w} \end{array}\right\}$$

which does not depend on $F(\cdot)$, where $\hat{\rho}_1^{\text{Poisson}}$, $\hat{\rho}_2^{\text{Poisson}}$ denote the expected utilizations of the primary and secondary class respectively in a partition with N_1^* primary agents who participate according to Poisson (n_∞). □

Proof of Lemma 4. (a) Conditioning that $\{\mathcal{N} = n\}$ agents participate where $\mathcal{N} \sim \text{Bin}(N, p)$, it is known that the expected waiting time $W(n)$ in the classic $M/M/n$ model is strictly decreasing in n. Hence, a coupling argument implies that the expected value $\mathbb{E}_{\mathcal{N}}[W(\mathcal{N})]$ is strictly decreasing in agents' participation probability p.

(b) Fix an arbitrary $k \in \{1,\ldots,N\}$ exogenously specified by the firm. Using order statistics notation, the expected total ability of the top k participating agents is given by

$$\sum_{i=1}^{k} \int_{a_{min}\left((\hat{\rho}_j)_{j=1}^{k}\right)}^{1} a\, dF_{N-i:\,N-1}(a)$$

which is maximized at the partition that minimizes a_{min}. That is, the optimal partition is the one that maximizes agents' participation probability and contains two priority classes as we show in Theorem 8. □

Proof of Theorem 10. Theorem 7 shows that $\lim_{N\to\infty} a_{min}^N = 1$. Let

$$P_j(a) := \left(\begin{array}{c} N-1 \\ j-1 \end{array} \right) F(a)^{N-j} (1 - F(a))^{j-1}$$

and note that for all $j \in \{1,\ldots,\overline{m}\}$ we have $0 \leq P_j(a) \leq P_j\left(a_{min}^N\right) \to \infty$. That is, $\lim_{N\to\infty} u\left(a_{min}^N\right) = \lim_{N\to\infty} \sum_{j=1}^{\overline{m}} \hat{\rho}_j \cdot P_j\left(a_{min}^N\right) = 0$ and the convergence is uniform. Since $u\left(a_{min}^N\right) \to 0 < c_p$, the unique threshold participation strategy of the agents has as a consequence that in the limit no agent participates as his expected utility does not cover his participation cost. The welfare optimization problem composed of firm's profit, the expected surplus of each agent net the expected waiting time of the customers is written as:

$$\max_{k,\,(\hat{\rho}_j)_{j=1}^{k}} \mathscr{W} = \underbrace{\Pi\left((\hat{\rho}_j)_{j=1}^{k}\right)}_{\text{firm's profit}} + \underbrace{\int_{a_{min}\left((\hat{\rho}_j)_{j=1}^{k}\right)}^{1} u(a)\, dF(a)}_{\text{participating agent expected surplus}} - \underbrace{\mathbb{E}_{\mathcal{N}}\left[W(\mathcal{N});\,(\hat{\rho}_j)_{j=1}^{k}\right]}_{\text{customers' expected waiting time}}$$

Note that the surplus of each individual agent with ability a is his expected utility $u(a)$. Since the ability of each agent is private information to him, $\int_{a_{min}}^{1} u(a)\, dF(a)$ reflects the expected surplus of a participating agent. The statement follows by combining Lemma 4 and Theorem 8 with the fact that total (participating) agent surplus goes to zero as $N \to \infty$. □

This page is intentionally left blank

Appendix $\mathcal{F}$

Proofs of Chapter 3

The following is a known result related to the exponential distribution that we use extensively in Chapter 3. We include its proof for concreteness.

Lemma 13. *Let $X_1, \ldots, X_N$ be independent random variables with X_i following an Exponential(μ_i) distribution. Then $\mathbb{P}\left[X_i = \min_{1\leq j\leq N}\{X_j\}\right] = \frac{\mu_i}{\sum_{j=1}^{N}\mu_j}$.*

Proof of Lemma 13. Conditioning we have

$$\begin{aligned}
\mathbb{P}\left[X_i = \min_{1\leq j\leq N}\{X_j\}\right] &= \int_0^\infty \mathbb{P}\left[X_i < X_j \text{ for } j\neq i |\ X_i = t\right] \mu_i e^{-\mu_i t} dt \\
&= \int_0^\infty \mathbb{P}\left[t < X_j \text{ for } j\neq i\right] \mu_i e^{-\mu_i t} dt \\
&= \int_0^\infty \prod_{j\neq i} \mathbb{P}\left[X_j > t\right] \mu_i e^{-\mu_i t} dt \\
&= \int_0^\infty \prod_{j\neq i} e^{-\mu_j t} \mu_i e^{-\mu_i t} dt \\
&= \mu_i \int_0^\infty e^{-(\mu_1+\ldots+\mu_N)t} dt \\
&= \frac{\mu_i}{\mu_1+\ldots+\mu_N}
\end{aligned}$$

which completes the statement. □

Proof of Lemma 5. The result follows by applying Wald's theorem to simplify (3.1). We show below that the assumptions of Wald's theorem are satisfied. Observe that the number of posted questions and the number of times the users and servers enter the forum have finite expectations since they belong to the finite interval $[0, T]$. Also, by assumption for each user $i = 1, \ldots, N$ (resp. for the servers) the arrival times $T_i(\mu_i p_i)$, $T_i(\mu_i(1-p_i))$ (resp. $T_f(s)$) into the forum are IID Exponential with rates $\mu_i p_i$, $\mu_i(1-p_i)$ (resp. s). Next, we have that the random variables $A_{q.}(\mu_i, p_i)$ are IID (they are independent by assumption, whereas they follow the same distribution due to the memoryless property of the exponential distribution applied to the arrival times of questions). Similarly, it follows that the random variables Q_e and Q_h are

IID. Hence, we have that

$$\begin{aligned} U_i(\mu_i, p_i) &= \mathbb{E}\left[\sum_{q_e \in Q_e}\left\{v_e \mathbb{1}_{A_{q_e}(\mu_i, p_i)}\right\} - \sum_{t \in T_i(\mu_i \cdot p_i)} c_e + \sum_{q_h \in Q_h}\left\{v_h \mathbb{1}_{A_{q_h}(\mu_i, p_i)}\right\} - \sum_{t \in T_i(\mu_i \cdot (1-p_i))} c_h\right] \\ &= \lambda_e T \cdot (v_e \mathbb{P}[A_{q_e}(\mu_i, p_i)]) + \lambda_h T \cdot (v_h \mathbb{P}[A_{q_h}(\mu_i, p_i)]) \\ &\quad - c_e \cdot (\mu_i \cdot p_i T + 1) - c_h \cdot (\mu_i \cdot (1 - p_i) T + 1), \end{aligned}$$

where we have used that $\mathbb{E}[Q_e(\mu_i, p_i)] = \lambda_e T$ (since this is average number of easy questions arriving in $[0, T]$) and $\mathbb{E}[T_i(\mu_i)] = \mu_i \cdot p_i T + 1$ (since each user enters into the forum one more time after T in case there are any unanswered questions and by assumption no more questions arrive after T).

Further, a user is rewarded for answering a given question if and only if he arrives before the servers and before the question expires. The users post answers to a given type of question at the rate they visit the forum multiplied by the probability they choose to answer that type of question. By Lemma 13 we have that $\mathbb{P}[A_{q_e}(\mu_i, p_i)] = \frac{p_i \cdot \mu_i}{p_i \cdot \mu_i + s + \theta}$. Hence, ignoring exogenous parameters user i solves

$$\max_{(\mu_i, p_i) \in [0, +\infty) \times [0,1]} \lambda_e v_e \frac{p \cdot \mu}{p \cdot \mu + s + \theta} - c_e p \cdot \mu + \lambda_h v_h \frac{(1-p) \cdot \mu}{(1-p) \cdot \mu + s + \theta} - c_h (1-p) \cdot \mu$$

as stated. □

Proof of Theorem 11. (i) Searching for a symmetric pure equilibrium consider the expected per question utility of a user

$$U(\mu, p) = \lambda_e v_e \frac{p \cdot \mu}{p \cdot \mu + s + \theta} - c_e p \cdot \mu + \lambda_h v_h \frac{(1-p) \cdot \mu}{(1-p) \cdot \mu + s + \theta} - c_h (1-p) \cdot \mu$$

Define $\mu_e := p \cdot \mu$ (resp. $\mu_h := (1-p) \cdot \mu$) be the rate of responding to easy (resp. hard questions) and we require that $\mu_e, \mu_h \geq 0$. Then, users' expected per question utility becomes

$$U(\mu_e, \mu_h) = \lambda_e v_e \frac{\mu_e}{\mu_e + s + \theta} + \lambda_h v_h \frac{\mu_h}{\mu_h + s + \theta} - c_e \mu_e - c_h \mu_h$$

Observe that the Hessian of U:

$$\mathcal{H}(\mu_e, \mu_h) = \begin{bmatrix} -\frac{2\lambda_e v_e (s+\theta)}{(s+\theta+\mu_e)^3} & 0 \\ 0 & -\frac{2\lambda_h v_h (s+\theta)}{(s+\theta+\mu_h)^3} \end{bmatrix}$$

is negative definite since all its eigenvalues are negative: $-\frac{2\lambda_e v_e (s+\theta)}{(s+\theta+\mu_e)^3} < 0$ and $-\frac{2\lambda_h v_h (s+\theta)}{(s+\theta+\mu_h)^3} < 0$ for every $\mu_e, \mu_h \geq 0$ (Sylvester's criterion). Hence, U is strictly concave.

The FOCs of $U(\mu_e, \mu_h)$ wrt μ_e and μ_h imply

$$\left.\begin{aligned} \frac{\lambda_e v_e (s+\theta)}{(s+\theta+\mu_e)^2} - c_e = 0 \\ \frac{\lambda_h v_h (s+\theta)}{(s+\theta+\mu_h)^2} - c_h = 0 \end{aligned}\right\}$$

Each equation gives two solutions of the form

$$\mu_{1,\cdot} = -\frac{1}{c.}\sqrt{c.\,\lambda.\,v.\,(s+\theta)} - (s+\theta) < 0$$

$$\mu_{2,\cdot} = \frac{1}{c.}\sqrt{c.\,\lambda.\,v.\,(s+\theta)} - (s+\theta) \geq 0 \tag{F.1}$$

The negative ones result in negative expected utility for the users ($U < 0$), that is the users do not participate in that case, i.e. they choose a rate $\mu^* = 0$ (since $U(\mu = 0) = 0$). Hence, we only keep the non-negative solutions:

$$\mu_e^* = \left(\frac{1}{c_e}\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta)} - (s+\theta)\right)^+, \quad \mu_h^* = \left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)} - (s+\theta)\right)^+,$$

where we used the notation $x^+ := \max\{x, 0\}$.

These imply the equilibrium (global) participation rate into the forum is

$$\mu^* = \mu_e^* + \mu_h^*$$

and conditional on participation (i.e. if $\mu^* > 0$) the equilibrium probability is

$$p^* = \frac{\mu_e^*}{\mu_e^* + \mu_h^*}$$

(if $\mu^* = 0$ the users do not participate and the equilibrium probability is undefined). Note that the equilibrium is unique and $p^* \in [0, 1]$, $\mu^* \geq 0$.

(ii) Each user arrives independently into the forum with rate μ^* and responds to easy questions with probability p^*. That is, we may think of users' responding to easy questions as performing N independent Bernoulli trials each having a "success" probability $\mu_e^* = \mu^* \cdot p^*$. For the hard questions the proof is similar. □

Proof of Theorem 12. (i) From Theorem 11 we have that the users' rate to easy (resp. hard) questions is $\mu_e^* = \left(\frac{1}{c_e}\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta)} - (s+\theta)\right)^+$ ($\mu_h^* = \left(\frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)} - (s+\theta)\right)^+$ respectively). The latter expressions become zero at $s_{0,e} := \left(\frac{\lambda_e\,v_e}{c_e} - \theta\right)^+$ (resp. $s_{0,h} := \left(\frac{\lambda_h\,v_h}{c_h} - \theta\right)^+$). Observe that our assumption $\frac{\lambda_e\,v_e}{c_e} < \frac{\lambda_h\,v_h}{c_h}$ implies that $s_{0,e} \leq s_{0,h}$ (with equality iff both are zero). If $s_{0,h} > 0$ then $\mu^*(s) = \mu_e^*(s) + \mu_h^*(s) > 0$ for each $s \in [0,\, s_{0,h})$ and $\mu^*(s) = 0$ for $s \geq s_{0,h}$. If $s_{0,h} = 0$ then $\mu^*(s) = 0$ and no user participates for all $s \geq 0$.

(ii) Observe that $\mu_e^*(s)$ and $\mu_h^*(s)$ are strictly concave wrt s with $\frac{d^2\mu_e^*}{ds^2}(s) = -\frac{c_e v_e^2}{4(c_e v_e(s+\theta))^{3/2}} < 0$ (similarly for $\frac{d^2\mu_h^*}{ds^2}(s)$). Hence, $\mu_e^*(s)$ and $\mu_h^*(s)$ have a unique maximum. The FOCs give the unique maxima $\frac{\lambda_e\,v_e}{4c_e}$ and $\frac{\lambda_h\,v_h}{4c_h}$ of $\mu_e^*(s)$ and $\mu_h^*(s)$ attained at $s_e^* := \left(\frac{\lambda_e\,v_e}{4c_e} - \theta\right)^+$ and $s_h^* := \left(\frac{\lambda_h\,v_h}{4c_h} - \theta\right)^+$respectively. From (i) if $s \geq \min\{s_{0,e},\, s_{0,h}\} = s_{0,e} > 0$ and $s_{0,e} < s_{0,h}$, we have that $\mu_e^*(s) = 0$ for $s \geq s_{0,e}$. Hence, the users always exploit and respond only to hard questions with rate $\mu_h^*(s) > 0$ when $s \in [s_{0,e},\, s_{0,h}]$.

Similarly, users' global service rate

$$\mu^*(s) = \mu_e^*(s) + \mu_h^*(s)$$

$$= \frac{1}{c_e}\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta)} + \frac{1}{c_h}\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta)} - 2\,(s+\theta)$$

is concave and hence it has a unique maximum. The FOC wrt s gives two solutions $s_1^* = \frac{\left(\sqrt{\lambda_e\,v_e\,c_h}-\sqrt{\lambda_h\,v_h\,c_e}\right)^2}{16c_ec_h} - \theta$ and $s^* = \frac{\left(\sqrt{\lambda_e\,v_e\,c_h}+\sqrt{\lambda_h\,v_h\,c_e}\right)^2}{16c_ec_h} - \theta$.

Observe that

$$\mu^*\,(s_1^*;\,\theta) = \frac{\lambda_h\,v_h\,c_e + 2\sqrt{c_e\,c_h\,\lambda_e\,\lambda_h\,v_e\,v_h} - 3c_h\,\lambda_e\,v_e}{8c_ec_h} < \mu^*\,(s^*;\,\theta) = \frac{\left(\sqrt{\lambda_e\,v_e\,c_h} + \sqrt{\lambda_h\,v_h\,c_e}\right)^2}{8c_ec_h}$$

Hence, $\mu^*\,(s)$ attains a unique maximum $\frac{\left(\sqrt{\lambda_e\,v_e\,c_h}+\sqrt{\lambda_h\,v_h\,c_e}\right)^2}{8c_ec_h}$ at $s^* = \frac{\left(\sqrt{\lambda_e\,v_e\,c_h}+\sqrt{\lambda_h\,v_h\,c_e}\right)^2}{16c_ec_h} - \theta$.

(iii) Observe that all else being equal, the firm's optimal rate obtained in Theorem 12(ii) $s^* = \frac{\left(\sqrt{\lambda_e\,v_e\,c_h}+\sqrt{\lambda_h\,v_h\,c_e}\right)^2}{16c_ec_h} - \theta$ is decreasing in c_e, c_h, and θ. We now show that this result continues to hold if we assume that the askers are heterogeneous in terms of their abandonment rate for each type of questions, i.e. $\theta_e \neq \theta_h$. If this is the case, we can not explicitly solve for s^* but instead we have that

$$\frac{d\mu^*}{ds}\,(s) = \frac{v_e\lambda_e}{2\sqrt{c_ev_e\lambda_e\,(s+\theta_e)}} + \frac{v_h\lambda_h}{2\sqrt{c_hv_h\lambda_h\,(s+\theta_h)}} - 2$$

Note that $\frac{d\mu^*}{ds}\,(s)$ is strictly decreasing in c_e and in θ_e. That is, $\mu^*\,(s)$ is strictly submodular in $(s,\,c_e)$ and in $(s,\,\theta_e)$ respectively, which immediately implies that the optimal s^* is strictly decreasing in c_e and in θ_e respectively. Similar arguments hold for c_h and in θ_h. □

Proof of Proposition 3. (i) Assume that the askers abandon service with rate θ_e (resp. θ_h) when posting easy (resp. hard) questions, and that these rates are different. Let $m\,(\theta_e,\,\theta_h) = \min\left\{\left(\frac{\lambda_e\,v_e}{c_e} - \theta_e\right)^+,\,\left(\frac{\lambda_h\,v_h}{c_h} - \theta_h\right)^+\right\}$ and $M\,(\theta_e,\,\theta_h) = \max\left\{\left(\frac{\lambda_e\,v_e}{c_e} - \theta_e\right)^+,\,\left(\frac{\lambda_h\,v_h}{c_h} - \theta_h\right)^+\right\}$. Arguing similarly to Theorem 12(i) we have that for $s \in [0,\,M\,(\theta_e,\,\theta_h))$ we have that $\mu^*\,(s) > 0$ and the users participate. Extending the definition of users' probability to resolve easy questions (Theorem 11) we have that $p^*\,(s;\,\theta_e,\,\theta_h) = \frac{\mu_e^*(s;\theta_e)}{\mu_e^*(s;\theta_e)+\mu_h^*(s;\theta_h)}$. Theorem 12 implies that one of the following holds: (Case 1) if $m\,(\theta_e,\,\theta_h) = \left(\frac{\lambda_e\,v_e}{c_e} - \theta_e\right)^+$, then $\mu_e^*\,(s) = 0$ for $s \geq \left(\frac{\lambda_e\,v_e}{c_e} - \theta_e\right)^+$ and $\mu_h^*\,(s) = 0$ for $s \geq \left(\frac{\lambda_h\,v_h}{c_h} - \theta_h\right)^+$, or (Case 2) if $m\,(\theta_e,\,\theta_h) = \left(\frac{\lambda_h\,v_h}{c_h} - \theta_h\right)^+$, then $\mu_h^*\,(s) = 0$ for $s \geq \left(\frac{\lambda_h\,v_h}{c_h} - \theta_h\right)^+$ and $\mu_e^*\,(s) = 0$ for $s \geq \left(\frac{\lambda_e\,v_e}{c_e} - \theta_e\right)^+$.

For both cases, the users respond to both types of problems, (i.e. $\mu_e^*\,(s)$, $\mu_h^*\,(s) > 0$ and $p^*\,(s;\,\theta_e,\,\theta_h) \in (0,\,1)$) if $s \in [0,\,m\,(\theta_e,\,\theta_h))$ (*exploration*). Under Case 1, the users only respond to hard questions (i.e. $p^*\,(s;\,\theta_e,\,\theta_h) = 0$) since $\mu_e^*\,(s) = 0$ and $\mu_h^*\,(s) > 0$ as long as $s \in [m\,(\theta_e,\,\theta_h),\,M\,(\theta_e,\,\theta_h))$ (*exploitation* of *hard* questions). Under Case 2, the users only respond to easy questions (i.e. $p^*\,(s;\,\theta_e,\,\theta_h) = 1$) since $\mu_h^*\,(s) = 0$ and $\mu_e^*\,(s) > 0$ as long as $s \in [m\,(\theta_e,\,\theta_h),\,M\,(\theta_e,\,\theta_h))$ (*exploitation* of *easy* questions).

(ii) From Proposition 3(i) we know that the users always exploit and respond only to either easy *or* hard questions with a positive rate w.p. 1 when $s \in [m\,(\theta_e,\,\theta_h),\,M\,(\theta_e,\,\theta_h))$, and do not participate for $s \geq M\,(\theta_e,\,\theta_h)$. Assume now that the servers set a rate $s \in [0,\,m\,(\theta_e,\,\theta_h))$.

Then,

$$p^*(c_e,\, c_h) = \frac{c_h\left(\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} - c_e\,(s+\theta_e)\right)}{c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} + c_e\left(-c_h\,(2s+\theta_e+\theta_h) + \sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)}\right)} \in (0,1)$$

with partial derivatives

$$\begin{aligned}\frac{\partial p^*}{\partial c_e}(c_e,\, c_h) &= \frac{c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)}\left(-\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)} + c_h\,(s+\theta_h)\right)}{2\left(c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} + c_e\left(-c_h\,(2s+\theta_e+\theta_h) + \sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)}\right)\right)^2} \\ &= -\mu_h^* \cdot \frac{c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)}}{2\left(c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} + c_e\left(-c_h\,(2s+\theta_e+\theta_h) + \sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)}\right)\right)^2} < 0\end{aligned}$$

and

$$\begin{aligned}\frac{\partial p^*}{\partial c_h}(c_e,\, c_h) &= \frac{c_e\sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)}\left(\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} - c_e\,(s+\theta_e)\right)}{2\left(c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} + c_e\left(-c_h\,(2s+\theta_e+\theta_h) + \sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)}\right)\right)^2} \\ &= \mu_e^* \cdot \frac{c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)}}{2\left(c_h\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta_e)} + c_e\left(-c_h\,(2s+\theta_e+\theta_h) + \sqrt{c_h\,\lambda_h\,v_h\,(s+\theta_h)}\right)\right)^2} > 0\end{aligned}$$

Observe that the sign of the above derivatives holds for all feasible values of the impatience thresholds θ_e and θ_h. All else equal, the equilibrium probability of choosing an easy question p^* is strictly decreasing in c_e and strictly increasing in c_h when $s \in [0,\, m(\theta_e,\, \theta_h))$ with $p^*(s) \in (0,1)$.

(iii) Assume that $c_e = c_h = c$ and $\theta_e = \theta_h = \theta$. Then, $m(\theta_e,\, \theta_h) = s_{0,\,e} = \left(\frac{\lambda_e\,v_e}{c_e} - \theta_e\right)^+$. For $s \in [0,\, s_{0,\,e})$ we have that

$$p^*(s) = \frac{\mu_e^*}{\mu_e^* + \mu_h^*} = \frac{c_h\left(\sqrt{c_e\,\lambda_e\,v_e\,(s+\theta)} - c_e\,(s+\theta)\right)}{\sqrt{c\,\lambda_e\,v_e\,(s+\theta)} + \sqrt{c\,\lambda_h\,v_h\,(s+\theta)} - 2c\,(s+\theta)}$$

with a strictly negative derivative

$$\frac{dp^*}{ds}(s) = \frac{c\left(\sqrt{c\,\lambda_e\,v_e\,(s+\theta)} - \sqrt{c\,\lambda_h\,v_h\,(s+\theta)}\right)}{2\left(-2c\,(s+\theta) + \sqrt{c\,\lambda_e\,v_e\,(s+\theta)} + \sqrt{c\,\lambda_h\,v_h\,(s+\theta)}\right)^2} < 0$$

Hence, $p^*(s)$ for $s \in [s_{0,\,e},\, s_{0,\,h}]$ is strictly decreasing wrt s. □

Proof of Proposition 4. From Theorem 11(ii) we have that the equilibrium number of user responses to easy (resp. hard) questions N_e (resp. N_h) follows $\text{Bin}(N,\, \mu_e^*)$ (resp. $\text{Bin}(N,\, \mu_h^*)$). It suffices to show that $\mu_e^*(s) < \mu_h^*(s)$ for each s, which would imply that $N_e < N_h$ almost surely (by a standard coupling argument). Since, $s_{0,\,e} = \frac{\lambda_e\,v_e}{c_e} - \theta < \frac{\lambda_h\,v_h}{c_h} - \theta = s_{0,\,h}$ we have that for $s \geq s_{0,\,e}$ $\mu_h^*(s) > 0$ and $\mu_e^*(s) = 0$.

Consider now the case $s \in [0,\, s_{0,\,e})$ where both rates are strictly positive, i.e. $\mu_h^*(s) > 0$ and $\mu_e^*(s) > 0$ for $s \in [0,\, s_{0,\,e})$. Then, $\mu_h^*(s) - \mu_e^*(s) = \sqrt{\frac{\lambda_h\,v_h\,(s+\theta)}{c_h}} - \sqrt{\frac{\lambda_e\,v_e\,(s+\theta)}{c_e}} > 0$. Hence, we have that $\mu_e^*(s) < \mu_h^*(s)$ for all $s \geq 0$. □

Proof of Lemma 6. Let μ^* and p^* be the users' participation rate and probability of responding

to easy questions in equilibrium. We note that the random variables $VC_e(s)$ and $VC_h(s)$ are IID, $\mathbb{E}[VC_{q_e}(s)] = \frac{N(p^*\mu^*)+s}{(N(p^*\mu^*)+s)+\theta}$ and $\mathbb{E}[VC_{q_h}(s)] = \frac{N\cdot(1-p^*)\mu^*+s}{(N\cdot(1-p^*)\mu^*+s)+\theta}$ (since both the N users *and* the servers respond to questions, and only answers arrived before the random patience time of the asker generate service value to the firm), and $\mathbb{E}[T_f(s)] = sT+1$ (since the servers enter into the forum one more time after T similarly to the users, while they do not enter thereafter as no more new content is generated). Arguing similarly to *Lemma 5* Wald's theorem implies

$$\begin{aligned}\Pi(s) &= \mathbb{E}\left[\sum_{q_e\in Q_e} V_e \mathbb{1}_{VC_{q_e}(s)} + \sum_{q_h\in Q_h} V_h \mathbb{1}_{VC_{q_h}(s)} - \sum_{t\in T_f(s)} c_f\right] \\ &= \lambda_e T\cdot(V_e\,\mathbb{P}[VC_{q_e}(s)]) + \lambda_h T\cdot(V_h\,\mathbb{P}[VC_{q_h}(s)]) \\ &\quad -(sT+1),\end{aligned}$$

Ignoring exogenous parameters, the servers solve

$$\max_{s\geq 0} \lambda_e V_e \frac{N\cdot(p^*\mu^*)+s}{(N\cdot(p^*\mu^*)+s)+\theta} + \lambda_h V_h \frac{N\cdot(1-p^*)\mu^*+s}{(N\cdot(1-p^*)\mu^*+s)+\theta} - c_f s$$

as stated. □

Proof of Theorem 13. Let $\mathcal{I}_1 := [0, s_{0,e}]$, $\mathcal{I}_2 := [s_{0,e}, s_{0,h}]$ and $\mathcal{I}_3 := [s_{0,h}, +\infty)$. From Theorem 12 and Theorem 11 we consider the following three cases. First, for a firm's rate $s\in\mathcal{I}_1$ the users respond to both easy and hard questions, hence in that region

$$\mu_e^* = p^*\cdot\mu^* = \frac{1}{c_e}\sqrt{c_e\lambda_e v_e(s+\theta)} - (s+\theta)$$

$$\mu_h^* = (1-p^*)\cdot\mu^* = \frac{1}{c_h}\sqrt{c_h\lambda_h v_h(s+\theta)} - (s+\theta)$$

and the firm's revenue becomes

$$R(s) = \frac{\lambda_e V_e N\cdot\left(\frac{1}{c_e}\sqrt{c_e\lambda_e v_e(s+\theta)}-(s+\theta)\right)+s}{\left(N\cdot\left(\frac{1}{c_e}\sqrt{c_e\lambda_e v_e(s+\theta)}-(s+\theta)\right)+s\right)+\theta} + \frac{\lambda_h V_h N\cdot\left(\frac{1}{c_h}\sqrt{c_h\lambda_h v_h(s+\theta)}-(s+\theta)\right)+s}{\left(N\cdot\left(\frac{1}{c_h}\sqrt{c_h\lambda_h v_h(s+\theta)}-(s+\theta)\right)+s\right)+\theta}$$

Second, when $s\in\mathcal{I}_2$ the users only respond to hard questions (i.e. $p^*=0$), and the firm responds to both easy and hard questions. Hence, when $s\in\mathcal{I}_2$ the firm's revenue becomes

$$R(s) = \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{N\cdot\left(\frac{1}{c_h}\sqrt{c_h\lambda_h v_h(s+\theta)}-(s+\theta)\right)+s}{\left(N\cdot\left(\frac{1}{c_h}\sqrt{c_h\lambda_h v_h(s+\theta)}-(s+\theta)\right)+s\right)+\theta}$$

Third, when $s\in\mathcal{I}_3$ no user participates into the forum, and only the firm responds to both easy and hard questions. That is, when $s\in\mathcal{I}_3$ the firm's revenue becomes

$$R(s) = \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{s}{s+\theta}$$

We next show that the firm's utility is strictly concave in s in each of the intervals $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$:

$$\Pi(s) = \lambda_e V_e \frac{N\cdot\mu_e^*(s)+s}{(N\,\mu_e^*(s)+s)+\theta} + \lambda_h V_h \frac{N\cdot\mu_h^*(s)+s}{(N\cdot\mu_h^*(s)+s)+\theta} - c_f s$$

Assume first that at least one of $\mu_e^*(s)$, $\mu_h^*(s)$ is strictly positive. Then

$$\frac{d^2\Pi}{ds^2}(s) = \lambda_e V_e \frac{-2\theta\left(1+N\frac{d\mu_e^*}{ds}(s)\right)^2 + N\theta\left(N\mu_e^*(s)+s+\theta\right)\frac{d^2\mu_e^*}{ds^2}(s)}{\left(N\mu_e^*(s)+s+\theta\right)^3} + \lambda_h V_h \frac{-2\theta\left(1+N\frac{d\mu_h^*}{ds}(s)\right)^2 + N\theta\left(N\mu_h^*(s)+s+\theta\right)\frac{d^2\mu_h^*}{ds^2}(s)}{\left(N\mu_h^*(s)+s+\theta\right)^3} < 0,$$

Indeed, the denominator of the first fraction is strictly positive since $\mu_e^*(s) \geq 0$, $s \geq 0$ and $\theta > 0$. Further, we have that

$$\frac{d^2\mu_e^*}{ds^2}(s) = \frac{-c_e v_e^2 \lambda_e^2}{4\left(c_e v_e\left(s+\theta\right)\lambda_e\right)^{\frac{3}{2}}} < 0$$

Similar arguments hold for the second fraction in firm's utility. If both $\mu_e^*(s)$, $\mu_h^*(s)$ are zero then

$$\Pi(s) = \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{s}{s+\theta} - c_f s$$

which is strictly concave as well with $\frac{d^2\Pi}{ds^2}(s) = -2\left(\lambda_e V_e \frac{\theta}{(s+\theta)^3} + \lambda_h V_h \frac{\theta}{(s+\theta)^3}\right) < 0$.

Since the firm's utility is strictly concave in the intervals $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$, it has a unique local maximum in each of them. Then, the unique global maximum of firm's utility is the maximum of these three maxima. □

Proof of Theorem 14. Suppose that users' population N is sufficiently large. We solve firm's problem given in Theorem 13 for a rate s belonging in each of the following three areas: $\mathcal{I}_1 = \left[0, \left(\frac{\lambda_e v_e}{c_e} - \theta\right)^+\right]$, $\mathcal{I}_2 = \left[\left(\frac{\lambda_e v_e}{c_e} - \theta\right)^+, \left(\frac{\lambda_h v_h}{c_h} - \theta\right)^+\right]$ and $\mathcal{I}_3 = \left[\left(\frac{\lambda_h v_h}{c_h} - \theta\right)^+, +\infty\right)$. Depending on the values of θ, $\frac{\lambda_e v_e}{c_e}$ and $\frac{\lambda_h v_h}{c_h}$ the first two intervals may empty. Thus, our proof has three parts, and for each part we find the local maxima of firm's utility in each of the non-empty intervals and compare them to find the global maximum, which is unique as argued in Theorem 13.

Part (A): Assume that $\frac{\lambda_e v_e}{c_e} - \theta > 0$. Since by assumption $\frac{\lambda_e v_e}{c_e} < \frac{\lambda_h v_h}{c_h}$, all intervals are non-empty.

Case 1: $s \in \mathcal{I}_1 = \left[0, \frac{\lambda_e v_e}{c_e} - \theta\right]$. In this region, firm's utility is given by

$$\Pi(s) = \lambda_e V_e \frac{N\cdot\mu_e^*(s)+s}{\left(N\mu_e^*(s)+s\right)+\theta} + \lambda_h V_h \frac{N\cdot\mu_h^*(s)+s}{\left(N\cdot\mu_h^*(s)+s\right)+\theta} - c_f s$$

with derivative

$$\frac{d\Pi}{ds}(s) = \frac{\lambda_e V_e \theta\cdot\left(N\cdot\frac{d\mu_e^*}{ds}(s)+1\right)}{\left(N\mu_e^*(s)+s+\theta\right)^2} + \frac{\lambda_h V_h \theta\cdot\left(N\cdot\frac{d\mu_h^*}{ds}(s)+1\right)}{\left(N\mu_h^*(s)+s+\theta\right)^2} - c_f \xrightarrow{N\to\infty} -c_f$$

For $s_1^* = 0$ we get

$$\begin{aligned}\Pi^*(0) &= \lambda_e V_e \frac{N \cdot \mu_e^*(0)}{N \mu_e^*(0) + \theta} + \lambda_h V_h \frac{N \cdot \mu_h^*(0)}{N \cdot \mu_h^*(0) + \theta} \\ &= \frac{\lambda_e V_e N \cdot \left(\frac{1}{c_e}\sqrt{c_e \lambda_e v_e \theta} - \theta\right)}{N \cdot \left(\frac{1}{c_e}\sqrt{c_e \lambda_e v_e \theta} - \theta\right) + \theta} + \frac{\lambda_h V_h N \cdot \left(\frac{1}{c_h}\sqrt{c_h \lambda_h v_h \theta} - \theta\right)}{N \cdot \left(\frac{1}{c_h}\sqrt{c_h \lambda_h v_h \theta} - \theta\right) + \theta} \\ &\xrightarrow{N\to\infty} \lambda_e V_e + \lambda_h V_h\end{aligned}$$

Hence, for sufficiently large N, $s_1^* = 0$ gives the local maximum $\Pi_1^* = \lambda_e V_e + \lambda_h V_h$, for $s \in \mathcal{I}_1 = \left[0, \frac{\lambda_e v_e}{c_e} - \theta\right]$.

Case 2: $s \in \mathcal{I}_2 = \left[\frac{\lambda_e v_e}{c_e} - \theta, \frac{\lambda_h v_h}{c_h} - \theta\right]$. In this region, firm's utility is given by

$$\Pi(s) = \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{N \cdot \mu_h^*(s) + s}{\left(N \cdot \mu_h^*(s) + s\right) + \theta} - c_f s$$

with derivative

$$\begin{aligned}\frac{d\Pi}{ds}(s) &= \frac{\lambda_e V_e \theta}{(s+\theta)^2} + \frac{\lambda_h V_h \theta \cdot \left(N \cdot \frac{d\mu_h^*}{ds}(s) + 1\right)}{\left(N \mu_h^*(s) + s + \theta\right)^2} - c_f \\ &\xrightarrow{N\to\infty} \frac{\lambda_e V_e \theta}{(s+\theta)^2} - c_f\end{aligned}$$

Asymptotically, the FOC for N large gives two solutions

$$s^* = \pm\sqrt{\frac{\lambda_e V_e \theta}{c_f}} - \theta$$

These are valid only if they belong in $\mathcal{I}_2$. If $\sqrt{\frac{\lambda_e V_e \theta}{c_f}} \le \frac{\lambda_e v_e}{c_e}$, then we set $s_{20}^* := \frac{\lambda_e v_e}{c_e} - \theta$. If $\sqrt{\frac{\lambda_e V_e \theta}{c_f}} - \theta \in \mathcal{I}_2$, we set $s_{21}^* := \sqrt{\frac{\lambda_e V_e \theta}{c_f}} - \theta$. Finally, if $\sqrt{\frac{\lambda_e V_e \theta}{c_f}} \ge \frac{\lambda_h v_h}{c_h}$ we set $s_{22}^* := \frac{\lambda_h v_h}{c_h} - \theta$. Observe that the local maximizer $s_{20}^* = \frac{\lambda_e v_e}{c_e} - \theta$ of $\mathcal{I}_2$ can never correspond to the global maximum of firm's utility. Indeed, firm's utility is decreasing in $\mathcal{I}_1$ and it is unimodal in $\mathcal{I}_2$, which implies that $s_{20}^* = \frac{\lambda_e v_e}{c_e} - \theta$ corresponds to a local minimum.

We have

$$\Pi^*(s^*) = \lambda_e V_e \frac{s^*}{s^* + \theta} + \lambda_h V_h \frac{N \cdot \mu_h^*(s^*) + s^*}{\left(N \cdot \mu_h^*(s^*) + s^*\right) + \theta} - c_f s^*,$$

which implies that

$$\lim_{N\to\infty} \Pi^*(s_{21}^*) = \lambda_e V_e + \lambda_h V_h + c_f \theta - 2\sqrt{c_f \theta \lambda_e V_e}$$

and

$$\lim_{N\to\infty} \Pi^*(s_{22}^*) = \lambda_e V_e + \lambda_h V_h + c_f \theta - c_h \frac{\lambda_e V_e \theta}{\lambda_h v_h} - c_f \frac{\lambda_h v_h}{c_h}$$

Hence, for sufficiently large N, we have the local maxima $\Pi_{21}^* = \lambda_e V_e + \lambda_h V_h + c_f \theta - 2\sqrt{c_f \theta \lambda_e V_e}$ and $\Pi_{22}^* = \lambda_e V_e + \lambda_h V_h + c_f \theta - c_f \frac{\lambda_e V_e \theta}{\lambda_h v_h} - \lambda_h v_h$, for $s \in \mathcal{I}_2 = \left[\frac{\lambda_e v_e}{c_e} - \theta, \frac{\lambda_h v_h}{c_h} - \theta\right]$.

Case 3: $s \in \mathcal{I}_3 = \left[\frac{\lambda_h\, v_h}{c_h} - \theta,\, +\infty\right)$. In this region, firm's utility is given by

$$\Pi\left(s\right) = \lambda_e\, V_e\, \frac{s}{s+\theta} + \lambda_h\, V_h\, \frac{s}{s+\theta} - c_f\, s$$

with derivative

$$\frac{d\Pi}{ds}\left(s\right) = \frac{\theta\, \left(\lambda_e\, V_e + \lambda_h\, V_h\right)}{\left(s+\theta\right)^2} - c_f$$

The FOC gives

$$s^* = \pm\sqrt{\frac{\theta\, \left(\lambda_e\, V_e + \lambda_h\, V_h\right)}{c_f}} - \theta,$$

which is valid as long as it belongs to $\mathcal{I}_3$. If $\sqrt{\frac{\theta\,(\lambda_e\, V_e + \lambda_h\, V_h)}{c_f}} \leq \frac{\lambda_h\, v_h}{c_h}$, then we set $s_3^* := \frac{\lambda_h\, v_h}{c_h} - \theta$. If $\sqrt{\frac{\theta\,(\lambda_e\, V_e + \lambda_h\, V_h)}{c_f}} > \frac{\lambda_h\, v_h}{c_h}$, we set $s_3^* := \sqrt{\frac{\theta\,(\lambda_e\, V_e + \lambda_h\, V_h)}{c_f}} - \theta$. Hence, we have another local maximum $\Pi_3^* = \lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - 2\sqrt{c_f\, \theta\, \left(\lambda_e\, V_e + \lambda_h\, V_h\right)}$, for $s \in \mathcal{I}_3 = \left[\frac{\lambda_h\, v_h}{c_h} - \theta,\, +\infty\right)$.

Since $\Pi_3^* < \Pi_{21}^*$ we have that for sufficiently large N there are three local maxima of firm's utility:

$$\Pi_1^* := \lambda_e\, V_e + \lambda_h\, V_h$$

$$\Pi_{21}^* := \lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - 2\sqrt{c_f\, \theta\, \lambda_e\, V_e}$$

$$\Pi_{22}^* := \lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - c_f\, \frac{\lambda_e\, V_e\, \theta}{\lambda_h\, v_h} - \lambda_h\, v_h$$

Observe that demanding s_{21}^* to belong to $\mathcal{I}_2$, i.e. $\frac{\lambda_e\, v_e}{c_e} - \theta \leq s_{21}^* \leq \frac{\lambda_h\, v_h}{c_h} - \theta$ or $\frac{\lambda_e\, v_e}{c_e} \leq \sqrt{\frac{\lambda_e\, V_e\, \theta}{c_f}} \leq \frac{\lambda_h\, v_h}{c_h}$ while simultaneously satisfying the assumptions of Part (A) and keeping $c_f\, \theta - 2\sqrt{c_f\, \theta\, \lambda_e\, V_e} > 0$ is not possible wrt c_f. Hence, Π_{21}^* is dominated by Π_1^*. Similarly, demanding $c_f\, \theta - c_f\, \frac{\lambda_e\, V_e\, \theta}{\lambda_h\, v_h} - \lambda_h\, v_h > 0$ and simultaneously satisfying the assumptions of Part (A) is not possible wrt c_f. Hence, Π_{22}^* is dominated by Π_1^*. Therefore, Π_1^* is the unique global maximum of firm's utility in Part (A) attained when firm does not interact in the forum at $s_1^* = 0$.

Part (B): Assume that $\frac{\lambda_e\, v_e}{c_e} - \theta \leq 0$ and $\frac{\lambda_h\, v_h}{c_h} - \theta > 0$. Here, $\mathcal{I}_1 := \emptyset$, and similarly to Part (A) comparing the local maxima of firm's utility when $s \in \mathcal{I}_2$ and when $s \in \mathcal{I}_3$ we have that $\Pi_3^* < \Pi_{21}^*$ so the global maximum of firm's utility is $\max\left\{\Pi_{21}^*,\, \Pi_{22}^*\right\}$ which lies in $\mathcal{I}_2$.

Part (C): Assume that $\frac{\lambda_h\, v_h}{c_h} - \theta \leq 0$. Here, $\mathcal{I}_1 := \emptyset$ and $\mathcal{I}_2 := \emptyset$ so the users do not participate and only the firm responds to questions. Similarly to Part (A) the global maximum of firm's utility is $\Pi_3^* = \lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - 2\sqrt{c_f\, \theta\, \left(\lambda_e\, V_e + \lambda_h\, V_h\right)}$ attained at $s_3^* = \left(\sqrt{\frac{\theta\,(\lambda_e\, V_e + \lambda_h\, V_h)}{c_f}} - \theta\right)^+$.

Overall, we have the following characterization of the global maximum of firm's problem:

$$\Pi^* = \begin{cases} \lambda_e\, V_e + \lambda_h\, V_h, & 0 < \theta < \frac{\lambda_e\, v_e}{c_e} \\ \max\left\{\lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - 2\sqrt{c_f\, \theta\, \lambda_e\, V_e},\quad \lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - c_f\, \frac{\lambda_e\, V_e\, \theta}{\lambda_h\, v_h} - \lambda_h\, v_h\right\}, & \frac{\lambda_e\, v_e}{c_e} \leq \theta < \frac{\lambda_h\, v_h}{c_h} \\ \lambda_e\, V_e + \lambda_h\, V_h + c_f\, \theta - 2\sqrt{c_f\, \theta\, \left(\lambda_e\, V_e + \lambda_h\, V_h\right)}, & \theta \geq \frac{\lambda_h\, v_h}{c_h} \end{cases}$$

attained at

$$s_1^* = 0,$$

$$s_{21}^{*} = \left(\sqrt{\frac{\lambda_e \, V_e \, \theta}{c_f}} - \theta \right)^{+} \text{ or } s_{22}^{*} = \left(\frac{\lambda_h \, v_h}{c_h} - \theta \right)^{+},$$

and

$$s_{3}^{*} = \left(\sqrt{\frac{\theta \, (\lambda_e \, V_e + \lambda_h \, V_h)}{c_f}} - \theta \right)^{+},$$

respectively. □

References

L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM, 2008.

P. Afèche and J. M. Pavlin. Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science*, 62(8):2412–2436, 2016.

L. Ales, S.-H. Cho, and E. Körpeoğlu. Optimal award scheme in innovation tournaments. *Operations Research, Forthcoming*, 2017.

A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Discovering value from community activity on focused question answering sites: a case study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 850–858. ACM, 2012.

A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, pages 95–106. ACM, 2013.

N. Archak and A. Sundararajan. Optimal design of crowdsourcing contests. *ICIS 2009 Proceedings*, 2009.

F. Baccelli, P. Boyer, and G. Hebuterne. Single-server queues with impatient customers. *Advances in Applied Probability*, pages 887–905, 1984.

S. Banerjee, C. Riquelme, and R. Johari. Pricing in ride-share platforms: A queueing-theoretic approach. *Available at SSRN 2568258*, 2015.

O. Baron, M. Hu, S. Najafi-Asadolahi, and Q. Qian. Newsvendor selling to loss-averse consumers with stochastic reference points. *Manufacturing & Service Operations Management*, 17(4):456–469, 2015.

K. J. Boudreau, N. Lacetera, and K. R. Lakhani. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, 57(5):843–863, 2011.

G. P. Cachon, K. M. Daniels, and R. Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management, forthcoming*, 2017.

H. Chesbrough. *Open business models: How to thrive in the new innovation landscape*. Harvard Business Press, 2013.

F. Cribari-Neto, N. L. Garcia, and K. L. Vasconcellos. A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2):269–277, 2000.

E. Dechenaux, D. Kovenock, and R. M. Sheremeta. A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669, 2015.

D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 119–128. ACM, 2009.

S. Erat and V. Krishnan. Managing delegated search over design spaces. *Management Science*, 58(3):606–623, 2012.

S. Erat and K. C. Lichtendahl Jr. Correction to Terwiesch and Xu (2008). *Unpublished Manuscript, Darden School of Business, University of Virginia*, 2015.

S. Erat and K. C. Lichtendahl Jr. Non-monotonic effort in optimal contests. *Unpublished Manuscript, Darden School of Business, University of Virginia*, 2016.

D. Fudenberg and D. K. Levine. Self-confirming equilibrium. *Econometrica*, 61:523–545, 1993a.

D. Fudenberg and D. K. Levine. Steady state learning and nash equilibrium. *Econometrica: Journal of the Econometric Society*, pages 547–573, 1993b.

A. Gaba and A. Kalra. Risk behavior in response to quotas and contests. *Marketing Science*, 18(3):417–434, 1999.

F. Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

N. Gans and Y.-P. Zhou. Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management*, 9(1):33–50, 2007.

A. Ghosh and J. Kleinberg. Incentivizing participation in online forums for education. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 525–542. ACM, 2013.

K. Girotra and S. Netessine. *The Risk-Driven Business Model: Four Questions That Will Define Your Company*. Harvard Business Press, 2014.

R. Gopalakrishnan, S. Doroudi, A. R. Ward, and A. Wierman. Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050, 2016.

I. Gurvich, M. Lariviere, and T. Moreno-Garcia. Operations in the on-demand economy: Staffing services with self-scheduling capacity. *Available at SSRN 2336514*, 2015.

I. Gurvich, M. Lariviere, and C. Ozkan. Coverage, coarseness and classification: Determinants of social efficiency in priority queues. *Available at SSRN 2857256*, 2016.

J. Hamari, M. Sjöklint, and A. Ukkonen. The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, 2015.

R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer, 2003.

B. Hu and S. Benjaafar. Partitioning of servers in queueing systems during rush hour. *Manufacturing & Service Operations Management*, 11(3):416–428, 2009.

R. Ibrahim. Managing queueing systems where capacity is random and customers are impatient. *Available at SSRN 2623502*, 2015.

S. Jain, Y. Chen, and D. C. Parkes. Designing incentives for online question-and-answer forums. *Games and Economic Behavior*, 86:458–474, 2014.

L. B. Jeppesen and L. Frederiksen. Why do users contribute to firm-hosted user communities? the case of computer-controlled music instruments. *Organization science*, 17(1):45–63, 2006.

A. Kalra and M. Shi. Designing optimal sales contests: A theoretical perspective. *Marketing Science*, 20(2): 170–193, 2001.

P. Kireyev. Markets for ideas: Prize structure, entry limits, and the design of ideation contests. *Working Paper, Harvard Business School*, 2016.

L. Kleinrock. Optimum bribing for queue position. *Operations Research*, 15(2):304–318, 1967.

C. Knessl. Some asymptotic results for the M/M/∞ queue with ranked servers. *Queueing Systems*, 47(3):201–250, 2004.

K. A. Konrad. *Strategy and dynamics in contests*. Oxford University Press, New York, 2009.

E. Körpeoğlu and S.-H. Cho. Incentives in contests with heterogeneous solvers. *Management Science, forthcoming*, 2017.

L. Kosten. Über Sperrungswahrscheinlichkeiten bei Staffelschaltungen. *Electra Nachrichten-Technik*, 14:5–12, 1937.

R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl. *Building successful online communities: Evidence-based social design*. MIT Press, 2012.

M. A. Lariviere and J. A. Van Mieghem. Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing & Service Operations Management*, 6(1):23–40, 2004.

S. Lenfle and C. Loch. Lost roots: how project management came to emphasize control over flexibility and novelty. *California Management Review*, 53(1):32–55, 2010.

K. C. Lichtendahl Jr, Y. Grushka-Cockayne, and P. E. Pfeifer. The wisdom of competitive crowds. *Operations Research*, 61(6):1383–1398, 2013.

T. X. Liu, J. Yang, L. A. Adamic, and Y. Chen. Crowdsourcing with all-pay auctions: a field experiment on taskcn. *Management Science*, 60(8):2020–2037, 2014.

LiveOps. The industry's most qualified talent in the cloud, 2014. [Accessed online on August 6, 2014, `http://goo.gl/q6tNSN`].

L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest Q&A site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 2857–2866. ACM, 2011.

R. Megidish and A. Sela. Allocation of prizes in contests with participation constraints. *Journal of Economics & Management Strategy*, 22(4):713–727, 2013.

J. Mihm and J. Schlapp. Sourcing innovation: Public and private feedback in contests. *Available at SSRN 2659171*, 2015.

B. Moldovanu and A. Sela. The optimal allocation of prizes in contests. *American Economic Review*, pages 542–558, 2001.

B. Moldovanu, A. Sela, and X. Shi. Contests for status. *Journal of Political Economy*, 115(2):338–363, 2007.

H. Nazerzadeh and R. S. Randhawa. Near-optimality of coarse service grades for customer differentiation in queueing systems. *Available at SSRN 2438300*, 2015.

S. Netessine and V. Yakubovich. The darwinian workplace. *Harvard Business Review*, 90(5):1–4, 2012.

G. F. Newell. *The M/M/∞ Service System with Ranked Servers in Heavy Traffic*. Springer-Verlag, 1984.

L. Nittala and V. Krishnan. Designing internal innovation contests. *Working Paper, Rady School of Management, University of California San Diego*, 2016.

O. Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.

S. O'Connor. Gig Economy: When your boss is an algorithm, 2016. [Accessed online on September 10, 2016, `https://goo.gl/QXUcA7`].

Z. J. Ren and Y.-P. Zhou. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383, 2008.

G. Roels and X. Su. Optimal design of social comparison effects: Setting reference groups and reference points. *Management Science*, 60(3):606–627, 2013.

M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer, 2007.

R. Siegel. All-pay contests. *Econometrica*, 77(1):71–92, 2009.

K. I. Stouras. Product support forums: Customers as partners in the service delivery. *Available at SSRN 2868382*, 2016.

K. I. Stouras, K. Girotra, and S. Netessine. LiveOps: The Contact Centre Reinvented. *INSEAD Case Study*, 2014.

K. I. Stouras, J. Hutchison-Krupat, and R. O. Chao. Motivating participation and effort in innovation contests. *Available at SSRN 2924224*, 2017.

K. I. Stouras, S. Netessine, and K. Girotra. First Ranked First To Serve: Strategic agents in a service contest. *Available at SSRN 2696868*, 2016.

X. Su and F. Zhang. Strategic customer behavior, commitment, and supply chain performance. *Management Science*, 54(10):1759–1773, 2008.

J. Surowiecki. *The Wisdom of Crowds*. Anchor Books, New York, 2005.

R. Swinney. Inventory pooling with strategic consumers: Operational and behavioral benefits. *Working Paper*, 2014.

C. S. Tang, J. Bai, K. C. So, X. M. Chen, and H. Wang. Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio. *Available at SSRN 2831794*, 2016.

T. Taylor. On-demand service platforms. *Available at SSRN 2722308*, 2016.

C. Terwiesch and K. T. Ulrich. *Innovation tournaments: Creating and selecting exceptional opportunities*. Harvard Business Press, 2009.

C. Terwiesch and Y. Xu. Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54(9):1529–1543, 2008.

E. von Hippel. *Democratizing Innovation*. MIT Press, 2005.

E. von Hippel. *Free Innovation*. MIT Press, 2016.

G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: an analysis of Quora. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1341–1352. ACM, 2013.

M. L. Weitzman. Optimal search for the best alternative. *Econometrica*, pages 641–654, 1979.

V. Yakubovich and D. Lup. Stages of the recruitment process and the referrer's performance effect. *Organization science*, 17(6):710–723, 2006.

H.-L. Yang and C.-Y. Lai. Motivations of Wikipedia content contributors. *Computers in human behavior*, 26(6): 1377–1383, 2010.

D. Zhan and A. R. Ward. Incentive based service system design: Staffing and compensation to trade off speed and quality. *Available at SSRN 2568007*, 2015.