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

Konstantinos I. Stouras

Darden School of Business, University of Virginia, Charlottesville, 22903 Virginia, kostas@virginia.edu

November 2016

Abstract. 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 askers' impatience, award and cost structure and types of questions attempted by the users towards characterizing how a firm should effectively manage such an innovative business model for service. Our results establish that the rate of firm's participation can steer users responses among the available question types, and that users' rate is unimodal in firm's rate up to a point where a very actively responding firm essentially discourages any participation of the users into its forum. Finally, from the practical perspective of managing customer support, our analysis offers rules of thumb on the extent to which firms should delegate service to a large online community of users, contingent on the incoming demand for service and available staffing level.

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

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 adopt *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[1] and Apple Support Communities[2] 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 paper 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.

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 close answered questions and also provide service to unresolved questions?

(2) How should a firm manage crowdsourcing service to its online community? How often 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 show that users' decision problem can be simplified into a per-question decision (Lemma 1). We then characterize users' equilibrium entry behavior, and equilibrium choice of questions to answer conditional on participation. We find that there are thresholds in firm's service rate that discourage any participation from the users for easy and hard questions respectively. Our analysis demonstrates that the rate of the users and the firm initially behave as complements, while for a sufficiently

[1] https://answers.microsoft.com/
[2] https://discussions.apple.com/

actively participating firm they become substitutes until a certain level where no user finds it rational to participate for any type of question.

Increasing firm's service rate initially motivates the users to respond to both type of questions with positive probability (Proposition 4). Despite any available high-cost-high-reward hard questions 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. 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 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 her 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 9 illustrated in Figure 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.

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 (2016) 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. Employing an all-pay simultaneous auction model DiPalantino and Vojnovic (2009) study users equilibrium choice among multiple auctions, 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. (2016) 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 make 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. 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

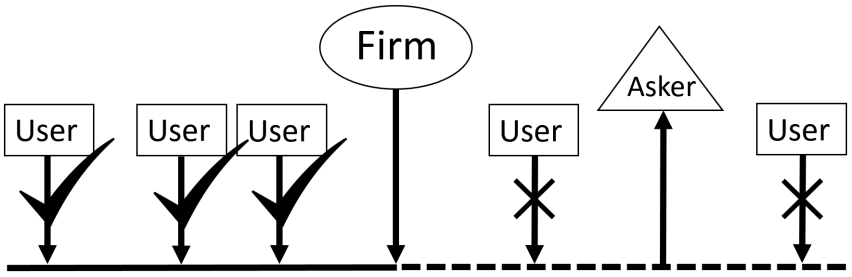

FIGURE 1: A product support forum model. All users that arrive before the firm and before the asker abandons service are rewarded by the asker; no other users are rewarded for the given question.

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 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 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

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

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[3] 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:

$$\Pi\left(s\right) = \mathbb{E}\left[\sum_{q_e \in Q_e} V_e \mathbb{1}_{VC_{q_e}(s)} + \sum_{q_h \in Q_h} V_h \mathbb{1}_{VC_{q_h}(s)} - \sum_{t \in T_f(s)} c_f\right], \tag{2}$$

where $VC_{q_\cdot}\left(s\right)$ 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 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 an 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

[3]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.

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.

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 1. *At a symmetric pure equilibrium each user solves*

$$\max_{(\mu,p)\in[0,+\infty)\times[0,1]} \lambda_e\, v_e \frac{p\cdot\mu}{p\cdot\mu+s+\theta} - c_e\, p\cdot\mu + \lambda_h\, v_h \frac{(1-p)\cdot\mu}{(1-p)\cdot\mu+s+\theta} - c_h\,(1-p)\cdot\mu \qquad (3)$$

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). 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 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, 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 2 characterizes the unique pure symmetric equilibrium of the service contest game played between the users and the firm.

Theorem 2 (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

$$\mu^* = \mu_e^* + \mu_h^* \qquad (4)$$

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

$$p^* = \frac{\mu_e^*}{\mu_e^* + \mu_h^*} \tag{5}$$

(ii) The equilibrium number of user responses to easy (resp. hard) questions follows Binomial(N, μ_e^*) (resp. Binomial(N, μ_h^*)).

Theorem 2(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 chose to participate. Conditional on participation, the users choose to respond to easy questions with probability p^* given by (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 4.

Further, a product support forum is a service system with random, on-demand capacity as in Stouras et al. (2016) and Ibrahim (2016). 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 2 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 3 (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 2 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.

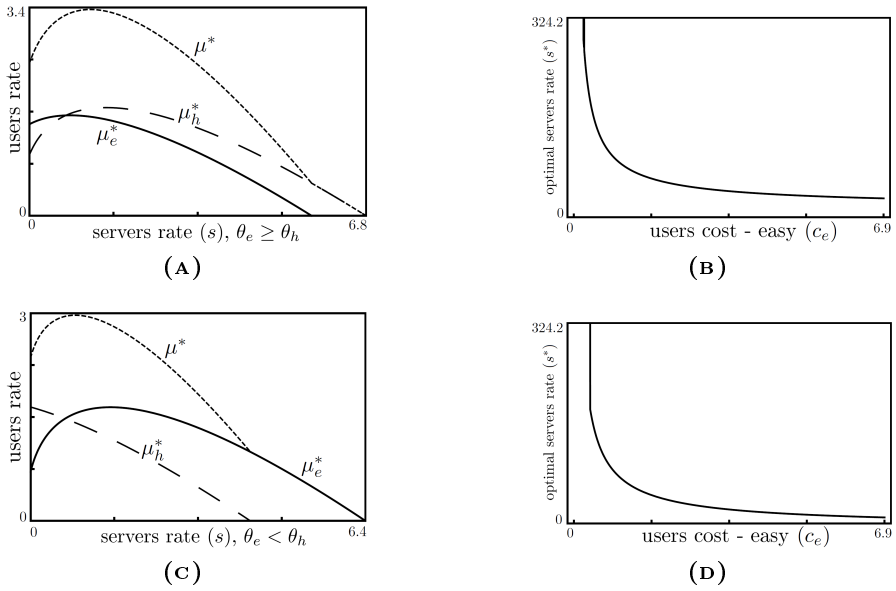

Figure 2: (A) and (C) Users' rate as a function of firm's rate; (B) and (D) Optimal firm's rate as a function of users' service rate cost for easy questions. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e,\, v_h,\, c_e,\, c_h) = (13,\, 35,\, 2,\, 5)$, $(\theta_e,\, \theta_h) = (0.8,\, 0.2)$ for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

(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.

A number of insights can be inferred from Theorem 3 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 $s_{0,h}$.

Theorem 3(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 3(ii) proves that users' rate on easy or hard questions is unimodal as a function of the firm's rate. In Figure 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 2(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 2(iii) illustrated in Figure 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 4 (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* $M(\theta_e, \theta_h) = \max\left\{\left(\frac{\lambda_e v_e}{c_e} - \theta_e\right)^+, \left(\frac{\lambda_h v_h}{c_h} - \theta_h\right)^+\right\}$.

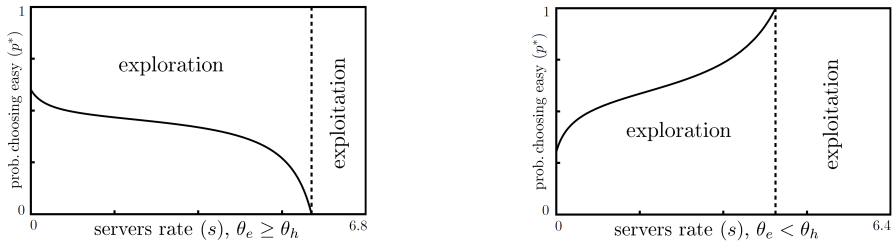

Figure 3: Users' equilibrium probability of replying to an easy question as a function of firm's rate. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, c_e, c_h) = (13, 35, 2, 5)$, $(\theta_e, \theta_h) = (0.8, 0.2)$ for (A), and $(\theta_e, \theta_h) = (0.1, 2.8)$ for (B).

(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 4(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 hard

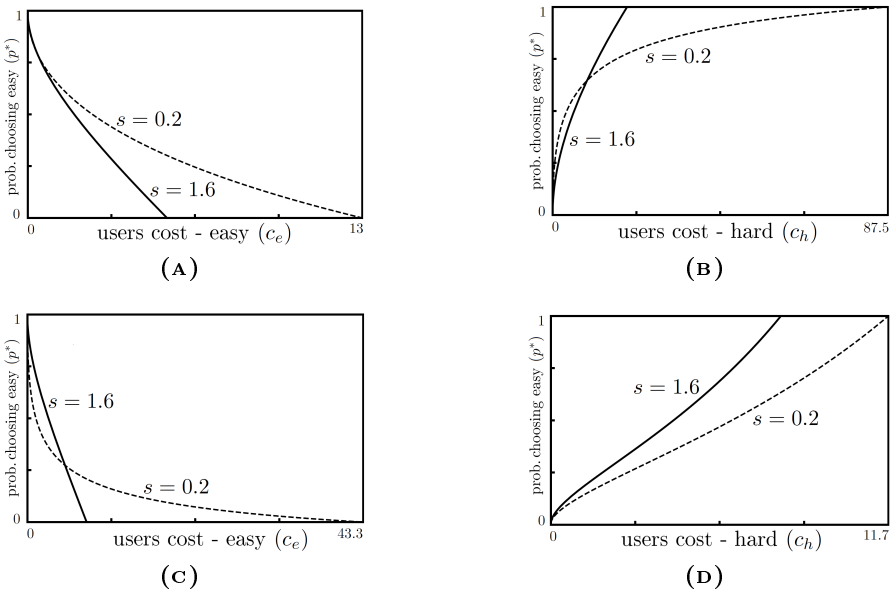

Figure 4: Users' equilibrium probability of replying to an easy question as a function of (A), (C) cost of easy questions, and (B), (D) cost of hard questions. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, c_e, c_h) = (13, 35, 2, 5)$, $(\theta_e, \theta_h) = (0.8, 0.2)$ for (A) and (B), and $(\theta_e, \theta_h) = (0.1, 2.8)$ for (C) and (D).

questions. As shown in Figure 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 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 become larger, users increasingly abandon service for easy questions searching for hard questions with higher potential. In light of the unimodality result of Theorem 2, 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 4(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 4(A) and Figure 4(B)).

In addition, when it is equally costly to reply to each type of questions Proposition 4(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 users who wish to cover their service rate and participation cost 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.

Proposition 5 (Clustering size)**.** *The number of user responses to hard questions is stochastically larger than the number of user responses to easy questions.*

Increasing participation rate in a service system 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 3 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 3.

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 §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.

5. Managing an online 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*

$$\max_{s\geq 0} \lambda_e V_e \frac{N\cdot(p^*\,\mu^*)+s}{(N\cdot(p^*\,\mu^*)+s)+\theta} + \lambda_h V_h \frac{N\cdot(1-p^*)\,\mu^*+s}{(N\cdot(1-p^*)\,\mu^*+s)+\theta} - c_f\, s \tag{6}$$

At a symmetric equilibrium firm's objective (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 7. *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 7 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

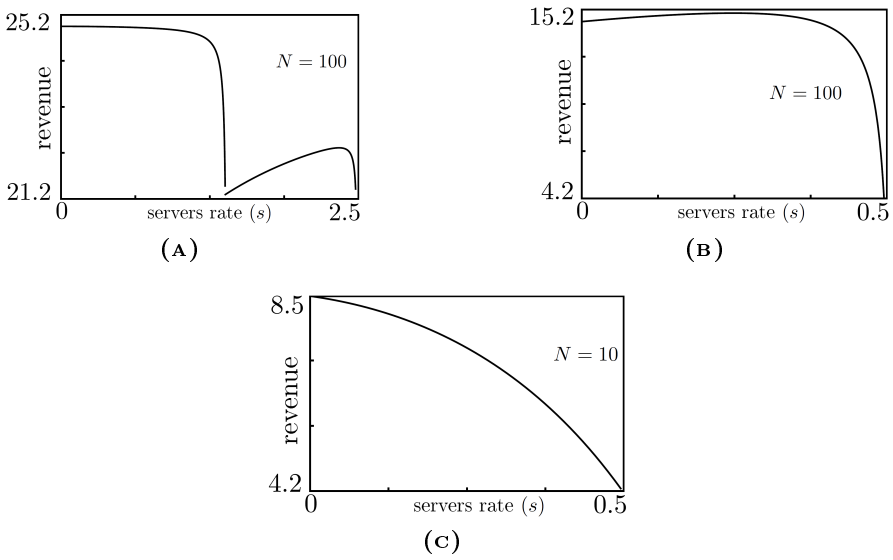

Figure 5: Firm's revenue as a function of servers' rate. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, V_e, V_h) = (13, 35, 10, 15)$ and $(c_e, c_h, c_f) = (6, 13.2, 2.2)$. Askers' impatience and users' population varies as follows: (A) $N = 100$ and $(\theta_e, \theta_h) = (0.8, 0.2)$; (B) $N = 100$ and $\theta_e = \theta_h = 2.2$; (C) $N = 10$ and $\theta_e = \theta_h = 2.2$.

question is solved), the following numerical examples (shown in Figure 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 the 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 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 5c).

We state the above observations as the following proposition.

Proposition 8. *In the presence of a participating online community, firm's revenue is non-monotone in firm's capacity.*

Theorem 7 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[4] in practice, we describe the optimal strategy for the firm in closed-form as follows.

Theorem 9. *Assume that there is a sufficiently large online community of users. Then, depending on askers' impatience level, the firm's utility is maximized at*

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

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

Theorem 9 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 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 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

[4]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).

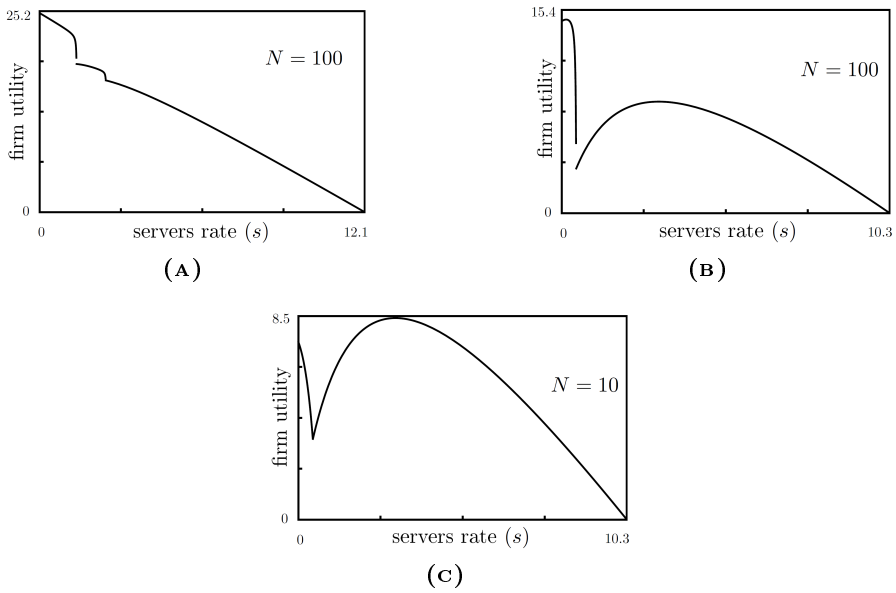

FIGURE 6: Firm's utility as a function of servers' rate. Parameters used: $\lambda_e = \lambda_h = 1$, $(v_e, v_h, V_e, V_h) = (13, 35, 10, 15)$ and $(c_e, c_h, c_f) = (6, 13.2, 2.2)$. Askers' impatience and users' population varies as follows: (A) $N = 100$ and $(\theta_e, \theta_h) = (0.8, 0.2)$; (B)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

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 6(b) with Figure 6(c)).

Theorem 9 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 9 on whether it is optimal for a firm to crowdsource its service operations in Figure 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 7, respectively), the effect of askers' impatience reflects what discussed in §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

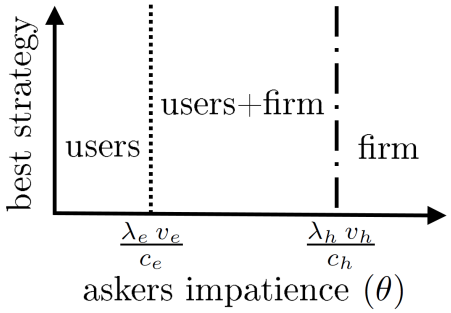

FIGURE 7: The optimal management of an online product support forum.

attract responses from the abundant users of the online community in order to alleviate the service congestion.

6. DISCUSSION

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 only hard 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 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 paper, 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.

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.

L. Ales, S.-H. Cho, and E. Körpeoglu. Optimal award scheme in innovation tournaments. *Forthcoming, Operations Research*, 2016.

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.

F. Baccelli, P. Boyer, and G. Hebuterne. Single-server queues with impatient customers. *Advances in Applied Probability*, pages 887–905, 1984.

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.

H. Chesbrough. *Open business models: How to thrive in the new innovation landscape*. Harvard Business Press, 2013.

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.

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.

R. Gopalakrishnan, S. Doroudi, A. R. Ward, and A. Wierman. Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050, 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. Ibrahim. Managing queueing systems where capacity is random and customers are impatient. *Working Paper*, 2016.

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.

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.

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.

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.

O. Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.

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.

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

C. Terwiesch and Y. Xu. Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54(9):1529–1543, 2008.

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.

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. Compensation and staffing to trade off speed and quality in large service systems. *Available at SSRN 2568007*, 2015.

Appendix A. Appendix

The following is a known result related to the exponential distribution that we use extensively in this paper. We include its proof for concreteness.

Lemma 10. *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 10. 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. □

Appendix B. Proofs

All proofs of the statements are given in their order of appearance in the main text.

Proof of Lemma 1. The result follows by applying Wald's theorem to simplify (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 Poisson 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. We have that

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

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 10 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 2. (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

$$\left.\begin{array}{l} \dfrac{\lambda_e\, v_e\ (s+\theta)}{(s+\theta+\mu_e)^2} - c_e = 0 \\ \dfrac{\lambda_h\, v_h\ (s+\theta)}{(s+\theta+\mu_h)^2} - c_h = 0 \end{array}\right\}$$

Each equation gives two solutions of the form

$$\mu_{1,\cdot} = -\frac{1}{c_\cdot}\sqrt{c_\cdot \lambda_\cdot v_\cdot\ (s+\theta)} - (s+\theta) < 0$$

$$\mu_{2,\cdot} = \frac{1}{c_\cdot}\sqrt{c_\cdot \lambda_\cdot v_\cdot\ (s+\theta)} - (s+\theta) \geq 0 \tag{7}$$

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 3. (i) From Theorem 2 we have that the users' rate to easy (resp. hard) questions is $\mu_e^* = \left(\frac{1}{c_e}\sqrt{c_e\, \lambda_e\, v_e\ (s+\theta)} - (s+\theta)\right)^+$ (resp. $\mu_h^* = \left(\frac{1}{c_h}\sqrt{c_h\, \lambda_h\, v_h\ (s+\theta)} - (s+\theta)\right)^+$). 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

$$\begin{aligned}\mu^*(s) &= \mu_e^*(s) + \mu_h^*(s) \\ &= \frac{1}{c_e}\sqrt{c_e \lambda_e v_e (s+\theta)} + \frac{1}{c_h}\sqrt{c_h \lambda_h v_h (s+\theta)} - 2(s+\theta)\end{aligned}$$

is concave and hence it has a unique maximum. The FOC wrt s gives two solutions $s_1^* = \frac{\left(\sqrt{\lambda_e v_e c_h} - \sqrt{\lambda_h v_h c_e}\right)^2}{16 c_e c_h} - \theta$ and $s^* = \frac{\left(\sqrt{\lambda_e v_e c_h} + \sqrt{\lambda_h v_h c_e}\right)^2}{16 c_e c_h} - \theta$. Observe that

$$\mu^*(s_1^*;\theta) = \frac{\lambda_h v_h c_e + 2\sqrt{c_e c_h \lambda_e \lambda_h v_e v_h} - 3 c_h \lambda_e v_e}{8 c_e c_h} < \mu^*(s^*;\theta) = \frac{\left(\sqrt{\lambda_e v_e c_h} + \sqrt{\lambda_h v_h c_e}\right)^2}{8 c_e c_h}$$

Hence, $\mu^*(s)$ attains a unique maximum $\frac{\left(\sqrt{\lambda_e v_e c_h} + \sqrt{\lambda_h v_h c_e}\right)^2}{8 c_e c_h}$ at $s^* = \frac{\left(\sqrt{\lambda_e v_e c_h} + \sqrt{\lambda_h v_h c_e}\right)^2}{16 c_e c_h} - \theta$.

(iii) Observe that all else being equal, the firm's optimal rate obtained in Theorem 3(ii) $s^* = \frac{\left(\sqrt{\lambda_e v_e c_h} + \sqrt{\lambda_h v_h c_e}\right)^2}{16 c_e c_h} - \theta$ 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

$$\frac{d\mu^*}{ds}(s) = \frac{v_e \lambda_e}{2\sqrt{c_e v_e \lambda_e (s+\theta_e)}} + \frac{v_h \lambda_h}{2\sqrt{c_h v_h \lambda_h (s+\theta_h)}} - 2$$

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, θ_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 4. (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 3(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 2) 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 3 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 4(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 5. From Theorem 2(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 1* 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 7. 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 3 and Theorem 2 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 $\mathcal{I}_3$:

$$\Pi(s) = \lambda_e\,V_e\,\frac{N \cdot \mu_e^*(s) + s}{(N\,\mu_e^*(s) + s) + \theta} + \lambda_h\,V_h\,\frac{N \cdot \mu_h^*(s) + s}{\left(N \cdot \mu_h^*(s) + s\right) + \theta} - c_f\,s$$

Assume first that at least one of $\mu_e^*(s)$, $\mu_h^*(s)$ is strictly positive. Then

$$\begin{aligned}\frac{d^2\Pi}{ds^2}(s) = \lambda_e\,V_e\,&\frac{-2\theta\,\left(1 + N\,\frac{d\mu_e^*}{ds}(s)\right)^2 + N\theta\,\left(N\,\mu_e^*(s) + s + \theta\right)\,\frac{d^2\mu_e^*}{ds^2}(s)}{\left(N\,\mu_e^*(s) + s + \theta\right)^3} \\ &+ \lambda_h\,V_h\,\frac{-2\theta\,\left(1 + N\,\frac{d\mu_h^*}{ds}(s)\right)^2 + N\theta\,\left(N\,\mu_h^*(s) + s + \theta\right)\,\frac{d^2\mu_h^*}{ds^2}(s)}{\left(N\,\mu_h^*(s) + s + \theta\right)^3} < 0,\end{aligned}$$

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

$$\frac{d^2\mu_e^*}{ds^2}(s) = \frac{-c_e\,v_e^2\,\lambda_e^2}{4\left(c_e\,v_e\;(s+\theta)\;\lambda_e\right)^{\frac{3}{2}}} < 0$$

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 9. Suppose that users' population N is sufficiently large. We solve firm's problem given in Theorem 7 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 7.

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

$$\Pi\left(s\right) = \lambda_e\, V_e\, \frac{N \cdot \mu_e^*\left(s\right) + s}{\left(N\, \mu_e^*\left(s\right) + s\right) + \theta} + \lambda_h\, V_h\, \frac{N \cdot \mu_h^*\left(s\right) + s}{\left(N \cdot \mu_h^*\left(s\right) + s\right) + \theta} - c_f\, s$$

with derivative

$$\begin{aligned}\frac{d\Pi}{ds}\left(s\right) &= \frac{\lambda_e\, V_e\, \theta \cdot \left(N \cdot \frac{d\mu_e^*}{ds}\left(s\right) + 1\right)}{\left(N\, \mu_e^*\left(s\right) + s + \theta\right)^2} + \frac{\lambda_h\, V_h\, \theta \cdot \left(N \cdot \frac{d\mu_h^*}{ds}\left(s\right) + 1\right)}{\left(N\, \mu_h^*\left(s\right) + s + \theta\right)^2} - c_f \\ &\xrightarrow{N\to\infty} -c_f\end{aligned}$$

For $s_1^* = 0$ we get

$$\begin{aligned}\Pi^*\left(0\right) &= \lambda_e\, V_e\, \frac{N \cdot \mu_e^*\left(0\right)}{N\, \mu_e^*\left(0\right) + \theta} + \lambda_h\, V_h\, \frac{N \cdot \mu_h^*\left(0\right)}{N \cdot \mu_h^*\left(0\right) + \theta} \\ &= \frac{\lambda_e\, V_e\, N \cdot \left(\frac{1}{c_e}\sqrt{c_e\, \lambda_e\, v_e\, \theta} - \theta\right)}{N \cdot \left(\frac{1}{c_e}\sqrt{c_e\, \lambda_e\, v_e\, \theta} - \theta\right) + \theta} + \frac{\lambda_h\, V_h\, N \cdot \left(\frac{1}{c_h}\sqrt{c_h\, \lambda_h\, v_h\, \theta} - \theta\right)}{N \cdot \left(\frac{1}{c_h}\sqrt{c_h\, \lambda_h\, v_h\, \theta} - \theta\right) + \theta} \\ &\xrightarrow{N\to\infty} \lambda_e\, V_e + \lambda_h\, V_h\end{aligned}$$

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

$$\Pi\left(s\right) = \lambda_e\, V_e\, \frac{s}{s + \theta} + \lambda_h\, V_h\, \frac{N \cdot \mu_h^*\left(s\right) + s}{\left(N \cdot \mu_h^*\left(s\right) + s\right) + \theta} - c_f\, s$$

with derivative

$$\frac{d\Pi}{ds}(s) = \frac{\lambda_e V_e \theta}{(s+\theta)^2} + \frac{\lambda_h V_h \theta \cdot \left(N \cdot \frac{d\mu_h^*}{ds}(s) + 1\right)}{\left(N \mu_h^*(s) + s + \theta\right)^2} - c_f$$
$$\xrightarrow{N\to\infty} \frac{\lambda_e V_e \theta}{(s+\theta)^2} - c_f$$

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 $\sqrt{\frac{\lambda_e V_e \theta}{c_f}} \leq \frac{\lambda_e v_e}{c_e}$, 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 $\sqrt{\frac{\lambda_e V_e \theta}{c_f}} \geq \frac{\lambda_h v_h}{c_h}$ 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

$$\Pi(s) = \lambda_e V_e \frac{s}{s+\theta} + \lambda_h V_h \frac{s}{s+\theta} - c_f s$$

with derivative

$$\frac{d\Pi}{ds}(s) = \frac{\theta\left(\lambda_e V_e + \lambda_h V_h\right)}{(s+\theta)^2} - c_f$$

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 $\max\{\Pi_{21}^*, \Pi_{22}^*\}$ 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 $\Pi_3^* = \lambda_e V_e + \lambda_h V_h + c_f \theta - 2\sqrt{c_f \theta \ (\lambda_e V_e + \lambda_h V_h)}$ 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:

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

attained at $s_1^* = 0$, $s_{21}^* = \left(\sqrt{\frac{\lambda_e V_e \theta}{c_f}} - \theta\right)^+$ 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. □