Operations Research Letters 42 (2014) 217–221

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Regulation under partial cooperation: The case of a queueing system

Gail Gilboa-Freedman *, Refael Hassin

Department of Statistics and Operation Research, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Received 10 December 2013
Received in revised form
13 February 2014
Accepted 25 February 2014
Available online 12 March 2014

Keywords:
Binary altruism
Partial control
Queueing

ABSTRACT

Socially optimal behavior can be achieved through the cooperation of the participating agents with a central planner. What happens when only a fraction of the population cooperates? We investigate this question in a Markovian single server queue. The main result is non-intuitive: the optimal control of cooperative customers is independent of their proportion. We also conclude that the gain from controlling cooperative customers *after* they join the queue is relatively small.

1. Introduction

Human altruistic behavior has been experimentally verified, and extensive literature analyzes its important consequences on economic behavior. The level of altruism is often measured as the degree of influence that the happiness or welfare of others has on the individual's own happiness or welfare. See, for example, [1,2,5] and their references. This literature considers agents that are *partially altruistic*. We consider, in contrast, a heterogeneous population where one's degree of altruistic behavior is dichotomous, being either totally selfish or totally devoted to the maximization of social welfare.

Our main motivation to investigate binary altruism comes from situations where a central planner's control of the population is limited, either because it can exercise authority on just a segment of the population, or because a request to cooperate with a certain policy is fulfilled by only a fraction of the population.

Examples are clearly ubiquitous: a demand to invest in clean energy can be enforced in some counties but not in others; a reduction in the use of a congested road can be exercised by controlling government vehicles but not private ones, or by tolling some roads but not others; only a fraction of the population will positively respond to the call to donate blood during an emergency or to be vaccinated to prevent an epidemic of an infectious disease, and not all drivers monitor travel information or advice from any one source—be it radio, TV, Internet, text messages, etc.; taxation and fines may be effective policy enforcing instruments in some segments of the population but may little change the behavior of others. In some situations, the choice of whether or not to comply with the central planner's instructions simply depends on the individual's degree of altruistic behavior (e.g., [14]). Staggered hours (e.g., [13]) and flextime (e.g., [19]) have been suggested in order to reduce road congestion during rush hours. A central question is whether, and how, a government can achieve a significant reduction of congestion by instructing those workplaces which are under its control to participate.

The fact that only a fraction of the population cooperates with the central planner's policy should obviously affect its instructions. For example, if pollution must be reduced to a given level, but cooperation is limited, then cooperative agents should be asked to reduce their emission of poisonous gases more than in the case of full cooperation. Two questions are naturally asked: What is the desired behavior of cooperative individuals given their number in the population, and how close to the socially maximal welfare can we get at a given level of cooperation.

For example, consider the following variation of the celebrated "tragedy of the commons": The size of the population, N, is very large. Individuals *independently* decide whether they participate in some activity, such that if X participate, then each of them obtains utility $N-X$. The expected value of the aggregate utility, $X(N-X)$, is maximized when each individual participates with probability 0.5, but when all are selfish, there is a unique equilibrium such that the whole population participates, with zero gain. Suppose now that a fraction α of the population is ready to cooperate. If $\alpha \leq 0.5$, then the social planner should instruct the cooperative individuals not to participate. If $\alpha > 0.5$ then they will be instructed to participate with probability $1-\frac{1}{2\alpha} < 0.5$. As intuitively expected, the lower the level of cooperation, the smaller the probability that a cooperative individual should participate. In other words, the

* Corresponding author.
E-mail addresses: email2gail@gmail.com, gailgf@post.tau.ac.il (G. Gilboa-Freedman), hassin@post.tau.ac.il (R. Hassin).

http://dx.doi.org/10.1016/j.orl.2014.02.006
0167-6377/

cooperative individuals compensate for the non-responsiveness of the non-cooperative fraction.

Of course, it is well known that the suboptimality of individual behavior as in this example, is caused because individuals ignore the negative externalities associated with participation. Investigating a general model of heterogeneous population may not lead to conclusive results, and we therefore focus on a certain structured setting. It turns out that this model is rich enough to provide unexpected insights.

We note that it is not always true that if only some agents adopt a socially beneficial policy then they should take "stronger" actions than if everyone participated. The theory of the second-best provides such examples (see, for example, [17,23]).

The seminal paper on a queueing system with strategic customers is Naor (1969) [20], see Hassin and Haviv (2003) [12] for an exposition of this literature. In Naor's model customers observe the queue length prior to their join-or-balk decision. The value of service is common and finite, and there is a constant waiting cost per unit time. Both the individual and social welfare maximizing strategies dictate joining iff the queue length is smaller than a given threshold. But the social welfare maximizing threshold is smaller than the equilibrium threshold.

Strategic queueing theory focuses on two types of situations. In one, customers are cooperative, obeying a central planner. In the other, customers are non-cooperative, maximizing their own welfare. In the real world, customers are diverse in their behavior. For example, legislation may force different disciplines on segments of the population.

We extend Naor's model by assuming a mixed population of controlled (or, cooperative) and uncontrolled (noncooperative) customers. A similar extension has been introduced [3], but, their model differs from ours in two crucial ways: the uncontrolled customers do not behave strategically (they always join the system), and the objective is to maximize the profit obtained from the controlled customers.

We consider two modes of control. With *admission control* the manager instructs cooperative customers to join or balk upon their arrival. With *dynamic control*, the manager can also instruct them to give up their position in the queue (without leaving the queue) at any time during their stay in the system, or even to abandon the queue and give up service. It is intuitively expected that the existence of noncooperative customers, should result in applying stricter admission rules on cooperative customers, because arrivals of selfish customers may generate long queues. This intuition is verified under admission control, but unexpectedly fails when dynamic control is exercised. This outcome is our main insight. The comparison of the two modes leads to the conclusion that the cost reduction obtained from dynamic control is in most cases quite small.

We also investigate the marginal effect of cooperation. We find that when the system is not heavily congested, cooperative customers have disproportional influence. When the system is congested, the influence of cooperation is greatest when the population has about the same number of cooperative and non-cooperative agents. This finding can be compared to [10].

The literature that has the highest resemblance to our model envisions a Stackelberg type game where the social planner leads by instructing the cooperative agents, and the non-cooperative agents follow. For examples, [15,16,6]. In contrast, decisions of the central planner in our model are made dynamically in response to the state of the system. The instructions of the central planner do not affect the equilibrium strategy of the non-cooperative customers.

There are examples in the queueing literature where only one of two customer classes can be controlled by the queue manager (for example, [18,9,22,4,7]), but in these models the two classes also differ by other attributes and the importance of controlling more customers is not a central issue. In our model, customers are homogeneous except for that some obey the central planner while the others are noncooperative.

2. The model

Naor (1969) considered a single server queue with Poisson arrival process of homogeneous customers with rate λ, and service of exponentially distributed duration with rate μ. (The assumptions of a single server and Poisson arrivals are not essential to our analysis, which can be extended to a G/M/s queueing system, as in [24].) The value of service is R per customer, and the cost of spending a unit time in the system is C. After normalization, the model has two parameters: the normalized traffic intensity $\rho = \frac{\lambda}{\mu}$ and the normalized value of service in terms of the expected waiting cost for a single service $\nu = \frac{R\mu}{C} \geq 1$.

The equilibrium solution follows a *pure threshold strategy*, namely, for some integer n_e, a customer joins the queue iff he observes at most $n_e - 1$ customers. It is straightforward that $n_e = \left\lfloor \frac{R\mu}{C} \right\rfloor = \lfloor \nu \rfloor$. The socially-optimal solution is also characterized by a threshold strategy n^*. Define $g(\nu) = \frac{\nu(1-\rho)-\rho(1-\rho^{\nu})}{(1-\rho)^2}$. Then $n^* = \lfloor \nu^* \rfloor$, where ν^* is the unique solution to $g(\nu) = \nu_e$. Naor observed that $n^* \leq n_e$.

We extend Naor's model by assuming that a proportion α of the customers are *cooperative* and obey a social planner, while the rest are *non-cooperative* and optimize their own welfare by following the threshold strategy n_e. We denote cooperative and non-cooperative customers by *c-customers* and *n-customers*, respectively.

3. Dynamic control

Dynamic control utilizes the discipline of c-customers to the maximal extent. The manager instructs c-customers to join or balk upon their arrival. Additionally, the manager can instruct them to renege or give up their position in the queue. In particular, since a c-customer can be instructed to leave at any time, there is no loss of optimality in giving priority to n-customers, and moreover when an n-customer arrives and a c-customer is served, service is preempted: the c-customer moves back to the queue and the arriving n-customer enters service. We assume that when a preempted service is resumed it starts from the point of preemption so that the discipline is work conserving. It follows that a strategy of the central planner is defined by the maximum number of c-customers $g(i)$ allowed to stay in the queue when there are *in*-customers in it.

The *naive strategy* is defined by $g(i) = \max(0, n^* - i)$. It is naive in the sense that c-customers behave as if all others are also cooperative. In this section we prove that the naive strategy is socially optimal.

A *transparent customer* (or, a *t-customer*) is a customer who obtains low priority relative to all other customers. In our discussion, there is at most one such customer. Thus, he is served only if there is no other customer in the system, and his service is preempted if a new customer arrives while the t-customer is in service. He is 'transparent' because the other customers are not affected by his existence in the system and therefore they ignore him.

Theorem 3.1. *Suppose that noncooperative customers behave according to individual thresholds all of which are at least n^*. Then, the naive strategy by the cooperative customers maximizes the expected social welfare of the system.*

Proof. Our proof follows the logic of Hassin [11]. Consider Naor's model, and a cooperative customer who arrives when the queue length is $n < n^*$. We refer to this customer as "tagged". Social optimality requires that the tagged customer joins the queue. Social welfare is not affected by the order customers are served, as long as the server is always busy when the system is not empty.