Economic Issues in Shared Infrastructures

Costas Courcoubetis and Richard Weber

***Abstract*—In designing and managing a shared infrastructure one must take account of the fact that its participants will make self-interested and strategic decisions about the resources that they are willing to contribute to it and/or the share of its cost that they are willing to bear. Taking proper account of the incentive issues that thereby arise, we design mechanisms which, by eliciting appropriate information from the participants, can obtain for them maximal social welfare, subject to charging payments that are sufficient to cover costs. We show that there are incentivizing roles to be played both by the payments that we ask from the participants and the specifiation of how resources are to be shared.**

New in this paper is our formulation of models for designing optimal management policies, our analysis that demonstrates the inadequacy of simple sharing policies, and our proposals for some better ones. We learn that simple policies may be far from optimal and that efficient policy design is not trivial. However, we find that optimal policies have simple forms in the limit as the number of participants becomes large.

***Index Terms*—Communication system economics, Grid computing, Incentives, Mechanism design, Scheduling, Virtualization.**

I. Introduction

INFRASTRUCTURE virtualization is a powerful tool towards the creation of a global computing and communication infrastructure. It allows organizations to cooperate and contribute physical resources to the creation of virtual facilities involving networking or computing and storage. Examples include virtual networks, computational grids and service clouds. Such facilties are shared by participating organizations and support specific services, applications or scientific experiments. Although virtualization technology has made significant progress, there remain many interesting and unanswered economic questions about the business models that can make such virtual infrastructures viable.

In this paper we make the fundamental assumption that each participant is an economic agent who profits from using the common infrastructure, but that the value which he places upon being allocated a quantity of resources is private information. His incentive is to obtain for himself as great as possible value from the shared infrastructure (or service), while contributing minimally to the costs of its formation and maintenance. The result is that the participants' individual aims are not aligned with overall system efficiency. This is an important observation and suggests that unless the appropriate incentives are in place, the economic performance of the resulting system may be greatly reduced. This raises the question of how a shared facility should be managed so as to resolve the unavoidable conflicts that arise between participants and to share the operating cost.

One way to share the cost of building a facility is to require participants to pay fees. Another way is to add together actual resources that participants contribute, the sum of which defines the size of the virtual facility. In this latter case, we say that the participants are making *payments in kind*. We might, in this case, operate a policy of asking each participant to choose for himself a quantity of resource that he will contribute to a shared pool of resources, and then say that at all future instants the resource pool will be shared amongst any participants who wish to draw upon it in proportion to the sizes of their contributions. The participant who contributes more will receive more. But might the system work better if the resource is shared in proportion to some other function of their contributions? It is questions like this that we address.

The problem of policy design for computing facilities is certainly not trivial, as has been observed in [1], [2], [3]. Simple policies may perform very badly if they do not incentivize participants truthfully to reveal privately-held information regarding the utility they will obtain from a given allocation of resources. There is recent work in [4], [5] concerning the definition of accounting requirements for grids, which suggests that more sophisticated policies can be implemented in practice.

In this paper we look at a number of models, making various assumptions about the parameters that can be measured, and discuss tools for defining optimal policies. These policies are designed to incentivize truthful revelation indirectly, by offering each participant a choice of options and then observing which of them he chooses. More specifically, agents' bids determine resource-sharing contracts. These contracts specify what quantities of resource each agent will obtain in each possible circumstance that some subset of agents wish to draw on the resource pool simultaneously. The parameters of the contracts become finalized once all agents have made their bids. Each participant is incentivized to bid truthfully, i.e. to reveal his true valuation for the given service. The resulting contracts provide optimal resource sharing, subject to a constraint that the fees paid by the agents cover costs. In this model the rules of running the system are defined as functions of the bids of the participants. We are effectively designing rules for a game (in which the agents play strategically) such that at the (Bayesian) Nash equilibrium the economic efficiency is maximized. Our approach leverages ideas from the theory for optimal auctions [6] and mechanism design [7] to the context of shared infrastructure design and management.

We must stress that the mathematics involved in construct-

C. Courcoubetis is with the Department of Computer Science, Athens University of Economics and Business, Athens 11362, GREECE, email: courcou@aueb.gr

R. Weber is with the Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB3 0WB, UK, email: rrw1@cam.ac.uk

Work of both authors has been partially supported by the FP7 project Gridecon.

ing optimal policies in the context of incomplete information can be very elaborate and rarely leads to simple analytic solutions. However, we can learn about some general features that good policies should have. So in sections that follow we look at a number of carefully chosen examples, many of which concern systems with just two participants. They all concern problems of efficiently sharing infrastructure amongst participants who hold private information. These examples are sufficiently simple that they can be solved, and their solutions point up important issues, challenges and future research directions, some of which we summarise in Section IX. We do make the critical assumption that agents know the value they place on being allocated resources, and also know the distributions of the private valuations of other agents. If we do not make those assumptions, then mechanism design problem trivializes or has not enough structure to lead to a solution. More work is needed to refine our results and investigate their translation to practical implementations, i.e. impact on job scheduling policies of existing systems, see [8] and [9].

Previous work on computational grids has recommended the formation of a market for computation and the use of prices at a heuristic level to guide resource sharing, see [10], [11], [12]. In this market providers (sellers) and consumers (buyers) of computing resources go to trade. In [13] an open market for trading computational resources is proposed, that operates similarly to the stock market double auction model, except that commodities are perishable. The market matches the asks and bids, just as in the stock market, and allocates resources accordingly. If it is competitive, then the market allocates resources efficiently, see [14], [15]. Organizations will decide how much infrastructure to self-procure and how much to obtain from the market based upon the equilibrium market price and on the statistics of their demand for computation, see [16]. This approach is sensible when resources are commoditized and the market is competitive. i.e. there is a large number of buyers and sellers for each resource type.

Our approach differs from the above, but is complementary and makes no assumption on competition. It is not based on a market; rather it *regulates* the system by setting rules to which participants must abide and a policy for sharing the resource pool and covering its cost. It is appropriate when organizations may collaborate over a long period of time either (i) to share the cost of running an existing facility, or (ii) to create a new shared virtual facility, by each contributing actual computing resources (or by providing finance for purchasing and maintaining those resources). Case (i) is common when the infrastructure is initially created using public funding, and (ii) is common in large e-science projects, e.g., [17], [18], [19], [20], and in other virtual facility building projects like OneLab [21] and PlanetLab [22]. Our approach allows for long-term predictable contracts, in which participants make contributions in kind (infrastructure). This can be preferred to the uncertainty of fluctuating prices in a dynamic market.

Mechanism design problems for scheduling have been considered by the computer science community, but with quite different objectives. In a problem addressed in [23] and [24] jobs must be allocated to machines which are strategic in revealing their processing times for the jobs. The aim is to find mechanisms that can compute (in polynomial time) both incentive payments and an allocation of jobs to machines, and obtain a makespan that is no more than some factor from optimal in the worst case. In addition to the emphasis on polynomial-time computation, there are two other differences to our work: (i) there is no notion of any a priori distributions, and so incentive compatibility conditions are to hold ex-post, and (ii) there is no constraint on the budget that is available to pay agents to reveal information (or conversely, a constraint that payments taken from agents must cover cost).

The paper is organized as follows. In Section II we introduce a model for sharing an infrastructure of a given size and then provide an example in Section III. Section IV considers a problem of determining the infrastructure size. The supporting theory for sections I–IV is in Section V. It is also applied to a scheduling a server in Section VI. In Section VII we discuss the inefficiencies of simple policies. Finally, in Section VIII, we look at how the design problem simplifies as the number of participants becomes large. We draw out some interesting lessons for practice in Section IX.

II. A model for infrastructure sharing

A. Sharing a given infrastructure

We begin by presenting a model for optimally sharing an infrastructure amongst n participating users (or agents). Our notion of an infrastructure is one that we deliberately keep quite general. It is composed of resources (such as links, servers and buffers) and it can be operated in various ways (by choice of scheduling, routing and the manner in which resources are shared by the agents). Let Ω denote the set of all such ways that the infrastructure can be operated.

To provide an example, let us introduce what we shall call the *scalar resource sharing model*. In this model the infrastructure can be parametrized by the quantity Q of a single resource. This might be the bandwidth of a communication link, or the cycles available in a computational grid. A manner of operation (i.e. a member ω of Ω) is specified by a vector $x_1, \dots, x_n$ denoting the allocations of the resource that are made to the agents, where $\sum_i x_i \leq Q$. The infrastructure is to be shared amongst agents on a sequence of days $1, 2, \dots$. There is daily operating cost, which we assume to be a constant c, for all $\omega \in \Omega$. This may include interest payments on capital investment.

Given that the infrastructure is operated at day t in manner $\omega \in \Omega$, the utility for agent i is $u_i(\omega|\theta_{i,t})$, where $\{\theta_{i,t}\}_{t=1,2,\dots}$ are independent and identically distibuted samples of a random variable with distribution function F_i. For convenience in exposition, let us suppose a product form $\theta_{i,t}u_i(\omega)$. The function $u_i(\cdot)$ is public knowledge, but $\theta_{i,t}$ is known only by agent i and is independent of other $\theta_{j,t}$, $j \neq i$. Essentially, one can think of $\{(u_1(\omega), \dots, u_n(\omega)) : \omega \in \Omega\}$, as a set of achievable points, whose values to the agents are uncertain to the operator of the facility because only the agents know $\theta_{1,t} \dots, \theta_{n,t}$. However, it is public knowledge, a priori, that $\theta_{i,t}$ is a sample of a random variable with distribution function F_i. In practice, the F_i might be constrained to a finite set of distributions, each associated with a certain organization type.

If $\theta_{i,t} = 0$ then agent i does not need to use the resource at day t.

We wish to operate the facility so as to maximize the total expected net benefit to the participating agents, subject to a constraint that they pay enough to cover a daily operating cost c. In doing this, we are to choose an *operating policy*, say M, implemented in two steps, as follows.

M1 (the rules) The agents are told that as a function of declared $\theta_t = (\theta_{1,t}, \dots, \theta_{n,t})$

(a) the operating policy in Ω will be chosen using the function $\omega(\cdot)$;

(b) the payments on day t will be determined by the function $p(\cdot) = (p_1(\cdot), \dots, p_n(\cdot))$.

M2 (the game) Knowing all data (that is, the F_j, u_j, for all $j \in \{1, \dots, n\}$), his own $\theta_{i,t}$, and the functions $\omega(\cdot)$ and $p(\cdot)$ in M1, and assuming all other agents are truthful in their declarations, agent i has a priori incentive to declare truthfully his $\theta_{i,t}$.

Given the declarations of $\theta_{1,t}, \dots, \theta_{n,t}$, steps (a) and (b) are now implemented.

In M1 the functions $\omega(\cdot)$ and $p(\cdot)$ define a game in which agents participate by declaring values for their privately known $\theta_{i,t}$. We wish to choose $\omega(\cdot)$ and $p(\cdot)$ so that at an equilibrium of this game some objective is achieved. For example, we might seek the greatest possible sum of agents' utilities. This makes our problem one of *mechanism design*. We now sketch, so far as space allows, those basic elements of formal mechanism design theory that are relevant. For more detail readers are referred to [7].

Let $(\eta, \theta_{-i,t})$ be shorthand for $(\theta_{1,t}, \dots, \eta, \dots, \theta_{n,t})$ and $E_{-\theta_{i,t}}$ denote expectation over all $\theta_{j,t}$ for which $j \neq i$. If agent i declares $\theta_{i,t} = \eta$ then his payoff is his *expected net benefit*

$$E_{-\theta_{i,t}}[\theta_i u_i(\omega(\eta, \theta_{-i,t})) - p_i(\eta, \theta_{{}_{i,t}})]. \tag{1}$$

Suppose this game has a *Bayesian Nash equilibrium* at which each agent declares the true value of his $\theta_{i,t}$, i.e. no agent can improve his expected net-benefit by unilaterally departing from a strategy of making truthful declarations. The term *Bayesian* refers to the fact that each agent calculates his expected net-benefit while knowing the F_j, the *distributions* of all other agents' $\theta_{j,t}$s, (and that they will declare these $\theta_{j,t}$s truthfully.

The *revelation principle* states that nothing is lost by restricting attention to mechanisms whose equilibria are such that all agents make truthful declarations (so-called *direct-mechanisms*). Doing so imposes an *incentive compatibility condition* that agent i should be made to reveal truthfully his privately-known $\theta_{i,t}$. This is what we are saying in M2, and in condition C1 of in Section II-C.

We may immediately distinguish two important cases. Let us recall that in welfare economics an allocation of resources is called *Pareto-efficient* if no agent's position can be improved without making some other agent's position worse. Any allocation that maximizes social welfare is Pareto-efficient. We measure an agent's position is measured by (1) and social welfare by (3) and (4), below.

A *first-best* (Pareto-efficient) allocation of resources is achieved when goods and services are traded in a perfectly competitive free market. The same efficiency can also be achieved by a central controller who has complete information about agent preferences and who has full centralized control. This can happen in what we shall call the *full information* case, i.e. when the operator of the system can somehow access the true values of the $\theta_{i,t}$s.

There are many ways that perfect competition can fail. One is if agents collude. Another is if agents have private information; in our models we call this the *partial information* case, i.e. the operator and agents know only a priori distributions of $\theta_{i,t}$s. Now one must be content with a *second-best* allocation of resources. The second-best is achieved by the system designer imposing rules (for a auction, or other mechanism design) so that when independent agents act strategically in their own self-interests, (in respect of actions and any privately held information that such actions may reveal) then within the resulting non-coopertative game, the equilibrium (or worst equilibrium if there are more than one) has the greatest possible efficiency.

Given any set of operating rules (not necessarily those for which second-best efficiency is obtained) the term *price of anarchy* is used for the quotient between the first-best social welfare and the social welfare that is obtained at the worst of the possible non-cooperative equilibria. Sometimes the price of anarchy tends to 1 as the number of participants becomes large. This happens in a bandwith sharing problem in which the heuristic control is an auction, [25], and in which consequently the quotient between first-best and second-best welfares also tends to 1. We see this also in Theorem 3 of Section VIII-B. However, in [26] we have considered a peer-to-peer file-sharing system and shown that a heuristic of a fixed participation fee is asymptotically as good as second-best, but that the price of anarchy remains bounded away from 1.

B. The full information case

In the full information case the best way to operate the system on day t would be by choosing $\omega \in \Omega$ as the maximizer of the *social welfare*, giving

$$\omega(\theta_t) = \arg\max_{\omega} \left\{ \sum_{i=1}^{n} \theta_{i,t} u_i(\omega) - c \right\}. \tag{2}$$

If the system is operated daily using $\omega(\theta_t)$, then as each day is statistically the same, the long run average socal welfare is

$$E_{\theta_t}\left[\sum_{i=1}^{n} \theta_{i,t} u(\omega(\theta_t)) - c \right]. \tag{3}$$

If (3) is negative then there is no way to run the system so that costs are covered. If (3) is nonnegative then one could ask from agent i a daily payment p_i that is less than his expected benefit of $E_{\theta_t}[\theta_{i,t} u(\omega(\theta_t))]$, also choosing $p_1, \dots, p_n$ so that $\sum_i p_i = c$. Then each agent gains positive net benefit and the total payments cover cost. The payment p_i need not be monetary. Instead, agent i could be asked to contribute a fixed quantity of virtual resources that is of value p_i, i.e. to make a *payment in kind*.

We can now also define an *infrastructure optimization problem*. Suppose that there is a possible space Θ of infrastructures, i.e. $\Omega \in \Theta$, each with a given cost $c(\Omega)$. The problem is to choose $\Omega \in \Theta$ that maximizes the social welfare

$$E_{\theta_t}\left[\sum_{i=1}^{n}\theta_{i,t}u_i(\omega(\Omega,\theta_t))-c(\Omega)\right], \tag{4}$$

where $\omega(\Omega,\theta_t)$ denotes the optimal operation of the specific infrastructure Ω.

To make things concrete, we now assume the scalar resource sharing model. The set Θ of possible infrastructures might be $\Theta=\{Q : Q\geq 0\}$. The daily cost of the facility is $c(Q)$. Suppose that Q is given. In the full information case, the optimal allocations are given by

$$x^*(\theta_t,Q)=\arg\max_{\sum_i x_i\leq Q}\left\{\sum_{i=1}^{n}\theta_{i,t}u_i(x_i)\right\}. \tag{5}$$

The infrastructure optimization problem is

$$\max_Q\left\{E\left[\sum_{i=1}^{n}\theta_{i,t}u(x_i^*(\theta_t,Q))\right]-c(Q)\right\}. \tag{6}$$

C. The partial information case

In practice, the $\theta_{i,t}$ are usually private information of the agents and they will act strategically when asked to reveal them. An agent might choose to declare an inaccurate value of $\theta_{i,t}$ in order to obtain a larger resource share. To incentivize truthful declarations the operator must introduce payments which depend on those declarations. Now agent i declares $\theta_{i,t}$ to maximize (1), his expected net benefit. This leads to the type of game described in Section II-A. At the Bayesian Nash equilibrium, we wish the following conditions to be satisfied.

C1. *Incentive compatibility*: Agents should find it in their interest to be truthful in declaring their $\theta_{i,t}$.

C2. *Participation (also called individual rationality)*: Agents should see positive net benefit from participation.

C3. *Cost coverage (also called budget-balance)*: Payments should cover the cost $c(Q)$.

C4. *Maximum expected social welfare (total net benefit)* is attained (subject to C1–C3).

Each of C1–C3 can be imposed in two senses. Consider agent i and let $\theta_{-i,t}=(\theta_{1,t},\dots,\theta_{i-1,t},\theta_{i+1,t},\dots,\theta_{n,t})$. The ex-ante (weak) sense means that for all $\theta_{i,t}$ the condition holds in expectation, before $\theta_{-i,t}$ is known to agent i (and assuming truthful declarations by all other agents). For example, for C2, in the scaler resource sharing model, this means

$$E_{\theta_{-i,t}}\big[\theta_{i,t}u(x_i(\theta_t))-p_i(\theta_t)\big]\geq 0. \tag{7}$$

The ex-post (strong) sense means that for all possible $\theta_{i,t},\theta_{-i,t}$ the condition holds. For C2, this means

$$\theta_{i,t}u(x_i(\theta_t))-p_i(\theta_t)\geq 0. \tag{8}$$

Similarly, ex-ante and ex-post versions of C3 are $E_\theta[p_1(\theta_t)+\cdots+p_n(\theta_t)]\geq c$ and $p_1(\theta_t)+\cdots+p_n(\theta_t)\geq c$.

Observe that the class of policies discussed so far is restricted to those that are memoryless. More generally, we could make the choice of ω, or the allocation of resources at time t, depend on a τ-length history of declarations up to time t, $\{\theta_{t-\tau+1},\dots,\theta_t\}$. The best policy of this type is surely very complicated to derive. We look at two extreme cases:

– *one-shot participation*: the facility runs for one day or forever, but each agent remains in the system for only one day (and so $\tau=1$);

– *long-term participation*: the facility runs forever and the same agents participate each day (so effectively $\tau=t$).

The ex-ante versions of C1–C3 are natural for models with infinite repetition, where by the law of large numbers the agents and the facility operator see time averages of profits and cost covering payments. The one-shot scenario differs if the facility runs only once because, as the operator does not see time averages, covering cost should be ex-post.

III. Example of sharing a fixed resource

A purpose of the examples in this section is to illustrate the more general results to be derived in Section V. To begin, let us take the scalar resource sharing model with $n=2$ agents and $Q=1$. We analyse memoryless mechanisms. On day t, agent i has utility $\theta_{i,t}u_i(x)$ for resource x, where $\theta_{i,1},\theta_{i,2},\dots$, are independent samples from $U[0,1]$.

A. The case $u_i(x)=x$

As we see in Section V, if $u_i(x)=x$ then optimal mechanisms allocate the resource (if at all) wholly to the agent declaring the greatest $\theta_{i,t}$. This makes the solution to our mechanism design problem equivalent to that of an optimal auction, and we can directly translate results. We now describe some possible policies. In what follows, we drop the suffix t from $\theta_{i,t}$, since we now think about a memoryless mechanism applied on a typical day.

The first-best policy: This allocates the full resource to the agent having maximum θ_i. Using the fact that the maximum of two independent random variables, each uniformly distributed on $[0,1]$, has mean $2/3$, we have

$$E\Big[\max_{x_1+x_2\leq 1}(\theta_1x_1+\theta_2x_2)\Big]=E\big[\max(\theta_1,\theta_2)\big]=2/3,$$

where expectations are with respect to θ_1,θ_2. So if agents are truthful about their θ_i without the mechanism needing incentivize this, then the expected social welfare is $2/3-c$.

What happens if we try to use the above sharing policy, but take no payments from agents, and so offer no incentives for them to be truthful? Clearly, every agent will declare $\theta_i=1$, and the operator might flip a coin to decide on the allocation. The social welfare becomes $1/2-c$, substantially less than the first-best of $2/3-c$.

Second-best mechanisms: Let us now examine some mechanisms that maximize social welfare under constraints C1–C3. The first mechanism satisfies all the constraints ex-ante. The second mechanism satisfies the cost covering constraint C3 in the stronger ex-post sense, but constraints C1–C2 ex-ante. The third mechanism is Vickrey auction type of mechanism that satisfies constraints C1–C2 ex-post and C3 ex-ante.

Mechanism 1: This operating rule of this mechanism can be seen as arising from (21) Section V-B and its payments from (23). The operating rule, M1(a), is that amongst those agents declaring $\theta_i \geq \bar{\theta}$ the resource is wholly allocated to the one who declares the greatest θ_i. If neither declares $\theta_i \geq \bar{\theta}$, then no resource is allocated. The value of $\bar{\theta}$ is a parameter of the mechanism. The payment rule, M1(b) is that if agent i declares θ_i then he is charged

$$p_i(\theta_i) = \tfrac{1}{2}(\theta_i^2 + \bar{\theta}^2)1_{\{\theta_i > \bar{\theta}\}}. \tag{9}$$

Let $z_i(\theta_1, \theta_2)$ be 1 or 0 as the item is or is not allocated to agent i, when the agents declare their parameters to be θ_1, θ_2. Agent 1, who is assuming that at equilibrium agent 2 is declaring truthfully, declares $\theta_1 = \eta$ to maximize his ex-ante net benefit of

$$E_{\theta_2}\left[\theta_1 z_1(\eta, \theta_2) - \tfrac{1}{2}(\eta^2 + \bar{\theta}^2)\right] = \theta_1\eta - \tfrac{1}{2}(\eta^2 + \bar{\theta}^2). \tag{10}$$

This incentives $\eta = \theta_1$ for $\theta_1 > \bar{\theta}$, so ex-ante C1 holds. The maximized ex-ante net benefit is $\frac{1}{2}(\theta_1^2 - \bar{\theta}^2) \geq 0$, so ex-ante C2 holds.

The value of $\bar{\theta}$ is chosen so that ex-ante C3 holds, i.e.

$$\begin{aligned} c = E_{\theta_1,\theta_2}[p_1(\theta_1) + p_2(\theta_2)] &= 2\int_{\bar{\theta}}^{1} \tfrac{1}{2}(w^2 + \bar{\theta}^2)\, dw \\ &= \tfrac{1}{3} + \bar{\theta}^2 - \tfrac{4}{3}\bar{\theta}^3. \end{aligned} \tag{11}$$

The right hand side increases from $1/3 = 0.33\dot{3}$ to a maximum of $5/12 = 0.416\dot{6}$, as $\bar{\theta}$ increases from 0 to $1/2$. Thus any cost can be covered, up to $0.416\dot{6}$.

In Figure 1 we plot the value of the expected social welfare as a function of c, and compare it to the first-best value. The qualitative lessons are that first-best and second-best coincide if c is small, but second-best is strictly worse if c is large. For very large c it is impossible to cover costs. It is easy to check that the mechanism does not achieve any of C1–C3 ex-post. As we now show with Mechanisms 2 and 3, it is possible to strengthen the ex-ante constraints to ex-post ones, either for C3, or for C1 and C2, but not for all. However, before turning to Mechanisms 2 and 3, it is interesting to compare Mechanism 1 to a heuristic policy H which might be used by a non-sophisticated facility operator. The social welfare obtained by H is also shown in Figure1.

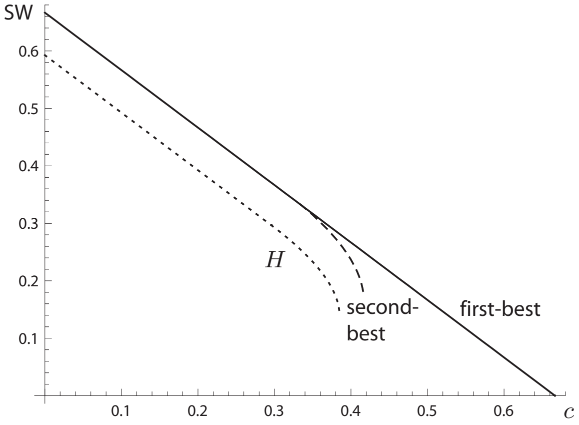

Fig. 1. Comparison of expected social welfares, as functions of c, for first-best (solid line), second-best (dashed line) and a heuristic H (dotted line). It is only for $c \in [0.333, 0.416]$ that the second-best falls short of the first-best. We see from (11) that there is no way to cover a cost $c \geq 0.416$ using a second-best mechanism.

A heuristic policy H: Suppose the operator posts a price p and ask both agents whether or not they are willing to pay this price in return for use of the resource. If just one is willing then he is allocated the resource and pays p. If both are willing then the resource is randomly allocated to one of them (by symmetry) and he pays p. Otherwise no resource is allocated and no payment is taken. The value of p is chosen so that the social welfare is maximized, subject to ex-ante covering of c. The best choice of p is found by solving the problem

$$\text{maximize } (1 - p^2)\tfrac{1}{2}(1 + p) \text{ s.t. } (1 - p^2)p \geq c.$$

It turns out that the optimal p is $p = 1/3$ if $c \leq 0.296$, and the smaller root of $(1 - p^2)p = c$, if $0.296 \leq c \leq 0.385$.

Mechanism 2: The allocation rule is as in Mechanism 1, but we adjust the payments so that C3 holds ex-post. This can be done by making agent 1 pay $p_1(\theta_1, \theta_2) = c/2 + p_1(\theta_1) - p_2(\theta_2)$, with $p_i(\theta_i)$ as defined in (9), i.e.

$$p_1(\theta_1, \theta_2) = \tfrac{1}{2}c + \tfrac{1}{2}(\theta_1^2 + \bar{\theta}^2)1_{\{\theta_1 > \bar{\theta}\}} - \tfrac{1}{2}(\theta_2^2 + \bar{\theta}^2)1_{\{\theta_2 > \bar{\theta}\}},$$

and similarly for agent 2. Note that C1 and C2 continue to hold ex-ante because $E_{\theta_2}[p_1(\theta_1, \theta_2)] = p_1(\theta_1)$ as in (9). Such an adjustment can always be made; if there were n agents one could take $p_i(\theta) = \frac{1}{n}c + p_i(\theta_i) - \frac{1}{n-1}\sum_{j \neq i} p_j(\theta_j)$.

Mechanism 3: The allocating rule is as in Mechanism 1. As an application of (24) agent 1 pays

$$p_1(\theta_1, \theta_2) = \max(\bar{\theta}, \theta_2)1_{\{\theta_1 > \max(\bar{\theta}, \theta_2)\}}$$

and similarly, agent 2. It gives a second-price (or Vickrey) auction. It is easy to check that C1–C2 hold ex-post. Moreover, $\bar{\theta}$ can be chosen so cost coverage is ex-ante, since it can be easily checked that $E_{\theta_2}p_1(\theta_1, \theta_2) = p_1(\theta_1)$ as in (9).

Mechanism 4 (for long-term participation): How might we exploit the fact of long-term participation? Might simpler mechanisms be more appropriate? Indeed this is true, as the following mechanism shows.

– At each time t, ask the agents to declare their $\theta_{i,t}$ and award the resource to the agent declaring greatest $\theta_{i,t}$.

– *Police the declarations*: make sure that in the long run the empirical distribution of the declared values of the $\theta_{i,t}$ matches F_i. If this is not the case then penalize agent i by imposing an appropriate charge.

– Assuming that agents are truth-telling, compute the expected benefit of each agent at each time t; then just as in our analysis of the full information case, split cost c arbitrarily into c_1 and c_2, so the expected benefit of agent i is at least c_i, and then use these fixed charges at each time t.

As we prove in Section V-A, this simple policy incentivizes truthful declarations. It also satisfies ex-ante C2 because in the long run each agent will have positive net benefit.

Our discussion to this point shows that operating policies may be very sensitive to modelling details. Policing is a practical option only when the same set of agents with known profiles are sharing the facility in a repeated fashion; this gives a special structure to the problem that allows us to achieve the same efficiency as in the full-information case. However, if agents change from day to day, or policing is not possible then we must to do something more interesting and nontrivial. We have seen that there can be several versions of second-best mechanisms; which we prefer can depend on which of constraints C1–C3 ought to be respected ex-post (or ex-ante) in given practical circumstances.

B. The case $u_i(x) = x^\beta$, $0 < \beta < 1$

With $u_i(x) = x^\beta$ ($0 < \beta < 1$) the story changes, in that the resource is no longer wholly allocated to the agent declaring greatest $\theta_{i,t}$. It turns out that under a second-best mechanism the optimal division of Q into x_1 and x_2 does not maximize $\theta_{1,t}x_1^\beta + \theta_{2,t}x_2^\beta$ (as it would be for first-best). There is an efficiency loss in inducing incentive compatibility. This will become clear in Section V-B.

Suppose that agents are identical and the common distribution of $\theta_{i,t}$ has density function f. Define, for $\lambda \geq 0$,

$$g(\theta_i) = \theta_i - \big(1 - F(\theta_i)\big)/f(\theta_i) \tag{12}$$

$$h_\lambda(\theta_i) = (\theta_i + \lambda g(\theta_i))^+. \tag{13}$$

We see in Section V that the optimal sharing policy is found by solving a Lagrangian dual problem:

$$\min_{\lambda \geq 0} \left\{ E\left[\max_{x_1+x_2 \leq 1} \sum_{i=1}^{2} h_\lambda(\theta_i) x_i^\beta \right] - (1+\lambda)c \right\}.$$

This means that $x_i(\theta_1, \theta_2) \propto h_\lambda(\theta_i)^{1/(1-\beta)}$. Notice that if $\lambda = 0$ this means allocating the resource in the most efficient way, i.e. to maximize $\sum_i \theta_i u(x_i)$. However, for a mechanism parameterized by $\lambda > 0$, the resource is allocated differently. There is a $\bar{\theta}$, such that an agent who declares $\theta_i < \bar{\theta}$ is allocated no resource. When $\theta_1 > \theta_2 > \bar{\theta}$, agent 1 receives a greater share of the resource than he would in an efficient allocation. We find $x_1(\theta_1, \theta_2)/x_2(\theta_1, \theta_2)$ is increasing in λ.

As λ increases from 0 to ∞ the cost that is being covered by the mechanism is increasing. For $\beta = 1/2$, this means we can cover $c \leq 0.2344$ with $\lambda = 0$, and then $c \in [0.2344, 0.4413]$ by taking $\lambda \in [0, \infty)$. The mechanism is complicated, but the results of its application can be calculated numerically and a figure produced that is very similar to Figure 1. We return to the issue of calculating payments for this mechanism in Section V-B. Similar stories are true if there are more than 2 agents. However, numerical calculations can be intractable.

IV. Building an optimal infrastructure

We return now to the infrastructure optimization problem that we touched upon briefly in Section II-B, in which the declarations of the agents are also used to choose the size of the infrastructure. There are now two stages in the implementation of a mechanism. In the first stage agents declare their 'types', that can be interpreted as how valuable is using the infrastructure on a *typical* day. As a function of these declarations the operator decides on the size of the infrastructure and on the payment functions and allocation rules that will be used in the second stage. In the second phase the infrastructure is shared among the agents according to the infinite repetition model introduced in Section II-A in which agents reveal their actual daily valuations.

Let us suppose a model in which if the infrastructure is operated at day t in manner $\omega \in \Omega$, the utility for agent i is $\phi_i \theta_{i,t} u(\omega)$, where ϕ_i is the *type* of the agent and $\theta_{i,t}$ is his normalized valuation of the service at the given day t. As before, it is public knowledge, a priori, that ϕ_i is a sample of a random variable with known distribution function Φ_i, but its true value is known only to agent i. Now, for each i, $\theta_{i,t}$ is also private information with a known distribution F_i. A way to interpret this is that $\theta_{i,t}$ states in a relative scale how valuable the service is to agent i, and multiplying it with ϕ_i rescales this to its actual value. For instance, $\theta_{i,t}$ could take a value in $\{1 \text{ (low)}, 2 \text{ (medium)}, 3 \text{ (high)}\}$, and multiplying it by ϕ_i leads to the actual service valuation. Or we could have that $\theta_{i,t}$ is uniform on $[0, 1]$, and ϕ_i is uniform on $[0, M]$, in which case $\phi_i \theta_{i,t}$ is uniform in $[0, \phi_i]$. Hence the type of agent i is needed in order to determine fully the distribution of the parameters $\theta_{i,t}$ in the model of Section II-A.

Our mechanism now is a modification of M that can be described as follows.

M†1 The agents are told that as a function of declared $\phi = (\phi_1, \ldots, \phi_n)$
  (a) the facility $\Omega(\phi) \in \Theta$ will be chosen;
  (b) the operating policy for declared θ_t will be $\omega(\theta_t, \phi)$;
  (c) the payments for declared θ_t will be $p(\theta_t, \phi) = (p_1(\theta_t, \phi), \ldots, p_n(\theta_t, \phi))$.

M†2 Knowing all the above, and assuming all other agents are truthful, C1–C3 hold at the equilibrium and agent i has a priori incentive to declare truthfully his ϕ_i, and subsequently at every day t his $\theta_{i,t}$.

Given declarations of $\phi_1, \ldots, \phi_n$, step (a) is implemented once at the start, and then, given $\theta_{1,t}, \ldots, \theta_{n,t}$, steps (b) and (c) are implemented daily.

This model of facility building has both one-shot and repeated components. The initial component is one-shot. On the basis of declared $\phi_1, \ldots, \phi_n$, the choice of Ω and the functions $\omega(\theta_t, \phi)$ and $p(\theta_t, \phi)$ are specified. Each agent (and the operator) can now estimate average net benefit (and revenue) throughout the future and might be inclined to leave the system (decline to operate) if this estimate is negative. This suggests that constraints C2–C3 must hold ex-post in regard to step (a). However, as regards to the repeated component of steps (b) and (c) they need only hold ex-ante. In general, it is impossible for all of C1–C3 to hold ex-post, but it becomes possible as the number if agents n tends to infinity.

Let us now describe a model that we shall often use. We use the terminology *activity model* to refer to modelling assumptions that the utility for agent i is $\phi_i \theta_{i,t} u(\cdot)$, where $\theta_{i,t} \in \{0, 1\}$, truthful declaration of $\theta_{i,t}$ takes place automatically, and ϕ_i is private information, with a priori distribution Φ_i. This gives a model in which agents are either 'active' or 'inactive', i.e. are only interested in using the infrastructure on some days, and given that agent i is active, she has always the same utility function $\phi_i u(\cdot)$. Assume that $\{\theta_{i,t}\}_{t=1,2,\ldots}$ are independent and identically distributed samples of a Bernoulli random variable. Let α_i be the *activity frequency* of agent i, i.e. the probability that $\theta_{i,t}=1$. Define the probability that on a given day the set of agents who wish to use the infrastructure S is $\alpha(S)$, where $S\subseteq\{1,\dots,n\}$. We assume that on any particular day the value of S is known, since there is no reason why any agent would pretend she wishes to use the resource when she cannot benefit from doing so, or pretend she cannot benefit from using it when she could. This is by the same arguments as in the proof of Theorem 1, Section V-A.

Consider the scalar resource sharing model, with the activity model assumptions, $n=2$, $u(x)=x$, and a priori $\phi_i\sim U[0,1]$. Suppose that $\Theta=\{Q:0\le Q\le 1\}$ and $c(Q)=\gamma Q$. Let $\bar\alpha_i=1-\alpha_i$. The first best optimum would be

$$\begin{aligned}&E\max_{0\le Q\le 1}\Big\{(\alpha_1\bar\alpha_2\phi_1+\alpha_2\bar\alpha_1\phi_2+\alpha_1\alpha_2\max(\phi_1,\phi_2))Q-\gamma Q\Big\}\\&=\Big(\alpha_1\bar\alpha_2\phi_1+\alpha_2\bar\alpha_1\phi_2+\alpha_1\alpha_2\max(\phi_1,\phi_2)-\gamma\Big)^{+}.\qquad(14)\end{aligned}$$

Recall that $h_\lambda(\phi)=\phi+\lambda g(\phi)$, where for $\phi_i\sim U[0,1]$ we have $g(\phi)=2\phi-1$.

As we see from the theory in Section V the value of the second-best social welfare is

$$\begin{aligned}&\min_{\lambda\ge 0}\Big\{E_{\phi_1.\phi_2}\Big[\Big(\alpha_1\bar\alpha_2 h_\lambda(\phi_1)+\alpha_2\bar\alpha_1 h_\lambda(\phi_2)\\&+\alpha_1\alpha_2\max\Big(h_\lambda(\phi_1),h_\lambda(\phi_2)\Big)-(1+\lambda)\gamma\Big)^{+}\Big]\Big\},\qquad(15)\end{aligned}$$

which also provides a way of finding the appropriate value of λ. This leads to a mechanism in which agent i participates only if $\phi_i>\bar\phi$ (defined by $h_\lambda(\bar\phi)=0$), and $Q=1$ or $Q=0$ as the term is round brackets above is positive or not.

V. Theory

This section contains the theory underlying Sections II–IV. We separately consider the scenarios in which a set of agents interacts over a long period of time (our ‘long-term participation’ model) or on just a single day (our ‘one-shot participation’ model). It is interesting that in the first of these scenarios a simple policy that uses policing can asymptotically obtain the same social welfare as the first-best policy, just as when there is full information.

A. Long-term participation: incentive compatibility achieved by policing

The key idea is that long-term participation makes it possible to incentivize agents to be truthful by policing their declarations. We may threaten to impose a very large fine upon agent i, or to exclude him from participation, if the empirical distribution of his declared $\theta_{i,1},\theta_{i,2},\dots$ does not converge to the publicly known F_i. There are many ways in which this might be done. For example, if F_i is the uniform distribution on $[0,1]$ we might partition $[0,1]$ into N equally likely subintervals of width $1/N$. We then run leaky bucket policers for each of these subintervals. Each bucket can hold an infinite number of tokens, and receives tokens at a rate of 1 per N days. A token is removed from the bucket corresponding to the subinterval of $[0,1]$ in which a declared $\theta_{i,t}$ falls; if there is no token in that buffer, then agent i obtains no resource. After many days (which should be exponential in N, so that the system will operates for a long time near its steady-state), N can be doubled. One can show (though we omit further details) that if this policer is employed then agent i maximizes his long-run average net benefit by respecting the constraint that the empirical distribution of his $\theta_{i,t}$ matches $U[0,1]$. Applying Theorem 1 (below) we can conclude that subject to this constraint he does best by being truthful.

Once we know that agents are truthful, the problem simplifies since we can then use (5) to make an optimal resource allocation for each vector $\theta_t=(\theta_{1,t},\dots,\theta_{n,t})$.

It remains to check that the combination of the allocation mechanism (5) and the policing mechanism described above does actually incentivize agents to be truthful. We need to check that there is no equilibrium which achieves a better payoff and in which agents sometimes report their $\theta_{i,t}$ in a non-truthful way. To check this, we start by noticing that the payment of p_i that is to be taken from agent i is fully determined by public knowledge of $F_1,\dots,F_n$, and so does not depend on the agent's declarations of $\theta_{i,1}$, $\theta_{i,2}$, … . Let us now consider whether it could be advantageous for agent i to decide that whenever his $\theta_{i,t}$ takes the value θ_i he will declare it to be θ_i' (possibly even randomizing). The policing mechanism constrains θ_i' to have the same distribution as θ_i. Subject to this constraint, the agent wishes to maximize $E[\theta_i V_i(\theta_i')]$, where $V_i(\theta_i)=E_{\theta_{-i}}[u(x_i(\theta_t))|\theta_{i,t}=\theta_i]$.

Theorem 1: Suppose that when agent i has $\theta_{i,t}=\theta_i$ he declares it as θ_i' (possibly randomizing), subject to the constraint that the unconditional distribution of θ_i' must also be F_i. Given that the resource is to be allocated according to declared $\theta_{i,t}'$ and by using (5), the agent maximizes his net benefit by always being truthful, i.e. with $\theta_i'=\theta_i$.

Proof: Given that $u_i(x)$ is concave increasing in x_i and $x_i(\theta_t)$ is determined by (5) the function $V_i(\cdot)$ must be nondecreasing. We now use the Hardy-Littlewood rearrangement inequality, which generalizes to integrals the simple fact that given any $a_1,\dots,a_n$ and $b_1,\dots,b_n$, then $\sum_i a_ib_i\le\sum_i a_i^*b_i^*$, where the starred sequences are rearrangements of the original sequences into increasing order. Taking $a(\theta_i)=\theta_i$ and $b(\theta_i)=E[V(\theta_i')|\theta_i]$, we have $a^*(\theta_i)=\theta_i$ and $b^*(\theta_i)=V(\theta_i)$, and obtain

$$\begin{aligned}E[\theta_iV_i(\theta_i')]&=E\Big[\theta_iE[V_i(\theta_i')|\theta_i]\Big]=\int_{\theta_i}ab\le\int_{\theta_i}a^*b^*\\&=E[\theta_iV_i(\theta_i)].\end{aligned}$$

So there is no reason for agent i to be other than truthful. ■

B. One-shot participation: optimal auctions

Now we turn to the more difficult circumstance in which it is not possible to police the parameters $\theta_{i,t}$ because the scenario is one-shot. Our discussion focuses upon a typical day t. Let $u_i(x_i|\theta_i)$ denote the utility of agent i for allocation of resource x_i. It is a function of his privately known parameter θ_i. A special case of this model is $u_i(x_i|\theta_i)=\theta_iu(x_i)$, as assumed hitherto. Another special case is

$$u_i(x|\theta_i)=\begin{cases}0, & x=0\\ r-(1-\theta_i)x, & x=1,2,\dots\end{cases}$$

for $0 < r \leq 1$. This models a scheduling problem, to be discussed in Section VI, in which there are n unit length jobs, each belonging to an agent; a subset of them us chosen for processing and scheduled in some order. If the job of agent i is completed after a time x, then he gains utility $r - \gamma_i x$, where $\gamma_i = 1 - \theta_i$ is a per unit cost of delay. The allocation $x = 0$ indicates that the job is not processed.

As usual, the a priori distribution of θ_i is F_i, which is known to all agents and the system operator. Based on this information, the operator imposes on the agents a mechanism, say M*. This is the same as M, except that we now allow the choice of operating mode ω to be a randomized choice within Ω, say taking the value ω with probability $q(\omega|\theta_t)$. The payments are $p_i(\omega, \theta_t)$, $i = 1, \dots, n$. Specification of the functions $q(\cdot|\cdot)$ and $p_i(\cdot, \cdot)$ are part of the rules of M*.

We now drop the suffix t, writing θ_i (and θ), in place of $\theta_{i,t}$ (and θ_t). For simplicity, let us suppose that for every pure choice of ω, the allocation $x_i(\omega)$ takes one of the values in a finite set, say X (e.g. $X = \{0, 1, \dots, Q\}$). Let us denote probabilities for agent i being allocated x, conditional on θ or on θ_i, as

$$\psi_i(x|\theta) = P(x_i(\omega) = x|\theta),$$
$$\psi_i(x|\theta_i) = P(x_i(\omega) = x|\theta_i) = E_{\theta_{-i}} \psi_i(x|\theta).$$

It is because M* allows randomization over the choice of ω that these variables can take values strictly between 0 and 1. We also denote ex-post and ex-ante payments as

$$p_i(\theta) = E_{\omega|\theta} p_i(\omega, \theta), \qquad p_i(\theta_i) = E_{\theta_{-i}} p_i(\theta).$$

The aim of the operator is to design a mechanism maximizing total expected net benefit of

$$E_\theta \left\{ \sum_i \sum_x u_i(x|\theta_i) \psi_i(x|\theta) \right\} - c, \tag{16}$$

subject to C1–C3. The ex-ante net benefit of agent i is

$$nb_i(\theta_i) = \sum_x u_i(x|\theta_i) \psi_i(x|\theta_i) - p_i(\theta_i). \tag{17}$$

For simplicity, suppose that $u_i(x|0) = 0$ for all x.

Theorem 2: There exists $\lambda \geq 0$ such that the optimal mechanism design (satisfying ex-ante C1–C3) chooses ω as function of θ to maximize

$$\sum_i \Big[(1 + \lambda) u_i(x_i|\theta_i) - \lambda \tfrac{1 - F_i(\theta_i)}{f_i(\theta_i)} \tfrac{\partial}{\partial \theta_i} u_i(x_i|\theta_i) \Big]. \tag{18}$$

Proof: The ex-ante expected net benefit of agent i, given in (17), is continuous and differentiable in θ_i. Assuming that the F_j are continuous, this is due to the averaging that takes place over θ_{-i} when obtaining $u_i(x|\theta_i) = E_{\theta_{-i}} u_i(x|\theta)$. The ex-ante incentive compatibility constraint C2 means that a truthful declaration of θ_i maximizes agent i's expected net benefit. Using (17), this provides a stationarity condition, that declaring $\theta_i = \eta$,

$$\sum_x u_i(x|\theta_i) \tfrac{\partial}{\partial \eta} \psi_i(x|\eta) - \tfrac{\partial}{\partial \eta} p_i(\eta) = 0.$$

Upon substituting $\eta = \theta_i$ and integrating, this gives

$$p_i(\theta_i) = p_i(0) + \sum_x \Bigg[u_i(x|\theta_i) \psi_i(x|\theta_i) - \int_0^{\theta_i} \tfrac{\partial}{\partial s_i} u_i(x|s_i) \psi_i(x|s_i) ds_i \Bigg]. \tag{19}$$

By taking an expected value of (19) with respect to θ, using integration by parts, and then summing on i, we find that the ex-ante cost-covering constraint can be written as

$$\sum_i p_i(0) + E_\theta \Bigg\{ \sum_i \sum_x \Big[u_i(x|\theta_i) - \tfrac{1 - F_i(\theta_i)}{f_i(\theta_i)} \tfrac{\partial}{\partial \theta_i} u_i(x|\theta_i) \Big] \psi_i(x|\theta) \Bigg\} \geq c.$$

Subject to this constraint, we wish to maximize (16). The decision variables are the $p_i(0)$ (which are to be ≤ 0) and the $\psi_i(x|\theta)$, $i \in \{1, \dots, n\}$, $x \in \{0, 1, \dots, Q\}$ (which are to be in $[0, 1]$, as well as consistent with the randomized choice of $\omega \in \Omega$). The fact that we allow the choice of ω to be randomized means that the set of all possible choices of decision variables is convex. All decision variables appear linearly in both the objective function and constraint, and so the problem can be solved by considering maximization of a Lagrangian of

$$\begin{aligned} L = E_\theta \Bigg\{ \sum_i \sum_x \Bigg[& (1 + \lambda) u_i(x|\theta_i) \\ & - \lambda \tfrac{1 - F_i(\theta_i)}{f_i(\theta_i)} \tfrac{\partial}{\partial \theta_i} u_i(x|\theta_i) \Bigg] \psi_i(x|\theta) \Bigg\} \\ & - (1 + \lambda) c + \lambda \sum_i p_i(0). \end{aligned} \tag{20}$$

This can be maximized pointwise for each θ, and the statement of the theorem now follows. ■

Let us make some remarks.

1. The assumption that $x_i(\omega)$ takes values in a finite set is simplifying for exposition, and useful in some examples. However, if x has a continuous domain we may replace $\sum_x \dots \psi_i(x|\theta)$ by $\int_x \dots \psi_i(x|\theta) dx$, with ψ now a density.

2. Theorem 2, its proof, and remark 3 recast, in the context of our models, standard arguments from the theory of optimal auctions and mechanism design, as expounded in [6] and [7]. There is one difference in that we seek to maximize social welfare, whereas in an auction one is usually seeking to maximize a principal's profit.

3. Consider the scalar resource sharing examples in Sections III-A and III-B, where we had $u_i(x_i|\theta_i) = \theta_i u(x_i)$. The coefficient of λ in (18) is $g_i(\theta) = \theta - (1 - F_i(\theta))/f_i(\theta)$. The choice of ω in (18) becomes the problem

$$\underset{\sum_i x_i \leq Q}{\text{maximize}} \left\{ \sum_i \big(\theta_i + \lambda g_i(\theta_i) \big) u(x_i) \right\}. \tag{21}$$

Assume $g_i(\cdot)$ is nondecreasing (as is the case for many distributions), and let $\bar{\theta}_i$ be the least θ_i for which it is profitable to allocate resource to agent i, i.e. $\theta_i + \lambda g_i(\theta_i) \geq 0$ for $\theta_i \geq \bar{\theta}_i$ in (21). To calculate the payments, define

$$V_1(w) = E_{\theta_2}\big[u(x_1(w, \theta_2))\big], \qquad (22)$$

and similarly $V_2(\cdot)$. Upon declaring θ_i agent i must pay

$$p_i(\theta_i) = \theta_i V_i(\theta_i) - \int_{\bar{\theta}_i}^{\theta_i} V_i(w) f_1(w) dw, \quad \theta_i \geq \bar{\theta}_i, \qquad (23)$$

and 0 otherwise.

This mechanism satisfies the ex-ante versions of C1–C3. It is possible, as above, to alter the mechanism so that ex-post version hold, either for C3, or for C1 and C2. For example, ex-post C1–C2 are achieved by agent 1 paying,

$$\begin{aligned} p_1(\theta_1, \theta_2) = \theta_1 u(x_1(\theta_1, \theta_2)) & \\ - \int_{\bar{\theta}_1}^{\theta_1} u(x_1(w, \theta_2)) f_1(w) dw, & \qquad (24) \end{aligned}$$

for $\theta_1 \geq \bar{\theta}_1$, and 0 otherwise, and similarly for $p_2(\theta_1, \theta_2)$.

Using L in (20), the appropriate value of λ can be found from the Lagrangian dual problem $\min_\lambda \max_x L$. We can write the maximum social welfare as

$$\min_{\lambda \geq 0} E\left[\max_{\sum_i x_i \leq Q} \left\{ \sum_i \big(\theta_i + \lambda g_i(\theta_i)\big) u(x_i) \right\} - (1+\lambda) c\right].$$

4. It is interesting to compare solutions with $\lambda > 0$ and $\lambda = 0$. For the following discussion, we continue to suppose that $u_i(x|\theta_i) = \theta_i u(x)$.

If $\lambda > 0$ then $E\big[\sum_i p_i(\theta_i)\big] = c$ and $p_i(0) = 0$ for all i. Note that the resource is not necessarily allocated in the same way that an efficient market would allocate it. For example, suppose $n = 2$ and $\theta_1 \sim U[0,1]$, $\theta_2 \sim U[0,2]$. Suppose that c is such that we cover the cost when taking $\lambda = 1$. Then if $\theta_1 = 5/6$ and $\theta_2 = 1$ we will have that x_1, x_2 should be chosen to maximize $\frac{3}{2} u(x_1) + u(x_2)$. Assuming u is concave this will mean we should take $x_1 > x_2$, even though $\theta_1 < \theta_2$.

If $\lambda = 0$ then we see from (18) that the resource is always allocated in the most efficient way, i.e. to maximize $\sum_i \theta_i u(x_i)$. This is now the same way an efficient market would allocate it. The expected sum of payments can create a surplus, say $s = E\big[\sum_i p_i(\theta_i)\big] - c > 0$. In this case we may take $p_1(0), \ldots, p_n(0)$ as any quantities summing to $-s$; for instance we could share the surplus equally amongst the agents by setting $p_i(0) = -s/n$.

5. At the end of Section IV we looked at an infrastructure optimization problem, in which the revelation of private parameters ϕ_1, ϕ_2 takes place once at the start and influences the choice of Q. The analysis for this problem is very similar. We derive (15) from the Lagrangian dual

$$\min_\lambda E_\phi \max_Q \Big\{ E_{\theta_t} \max_{x \in X} \textstyle\sum_i h_\lambda(\phi_i) \theta_{i,t} u(x_i) - (1+\lambda) c(Q) \Big\}.$$

This illustrates that one needs to be careful in ordering operators of $\max_Q$, E_ϕ and E_θ.

6. Although the above gives a methodology, it is not easy to apply analytically, even in simple cases. It is not even easy to say whether or not $\lambda = 0$, although we know this depends on the value of c.

VI. An application to scheduling a server

Suppose that each of n agents has a single unit length job which he wishes to have processed, and with minimum delay cost. An operator owns a machine. He is to decide which of the agents' jobs to process and how their processing is to be ordered. Suppose that the utility to agent i if his job is finished after a time x_i is $u_i(x_i) = r_i - \gamma_i x_i$, where $\gamma_i = 1 - \theta_i$ is the per unit time delay cost and θ_i is private information of agent i. If his job is not processed utility is 0. To indicate that a job is not processed we can let $x_i = 0$, with $u_i(0) = 0$. Suppose that a priori θ_i is distributed uniformly on $[0,1]$ and the r_is are known. Since jobs are of unit length, $x_i = j$ if agent i's job is processed jth in the sequence. The operator wishes to maximize the expected sum of net benefits, subject to obtaining payments from the agents sufficient to cover the cost of operating the machine, c. Application of our theory in Section V reveals that the optimal schedule maximizes, for some appropriately chosen λ,

$$\sum_i \Big[(1+\lambda)(r_i - (1-\theta_i)x_i) - \lambda \theta_i x_i\Big] 1_{\{x_i > 0\}}$$

$$= \sum_{i \in S} \Big((1+\lambda) r_i + \theta_i x_i\Big) - (1+\lambda) \sum_{i=1}^{|S|} i$$

where S is the set jobs chosen for processing. Consider the special case that $r_i = r$ for all i ($r < 1$), and suppose declarations are such that $\theta_1 > \cdots > \theta_n$. Then the mechanism will operate by choosing some set of jobs $1, 2, \ldots, k$ (with least delay costs) and then schedule them in order $k, \ldots, 2, 1$ (i.e. giving decreasing priority to jobs with decreasing delay costs). A little algebra shows that k is the least nonnegative integer such that

$$\frac{\theta_1 + \cdots + \theta_{k+1}}{k+1} < (1+\lambda)\left(1 - \frac{r}{k+1}\right), \qquad (25)$$

or $k = n$ if the above does not hold for $k = n$. Thus we have found the general form of an optimal operating policy. One might have guessed that an optimal mechanism would choose to process a set of jobs with small delay cost, but the precise criterion for selection in (25) is not something that one would easily guess. However, there remains a difficult calculation to determine the right payments, $p_i(\theta_i)$, and to find the value of λ such that the resulting policy induces payments that exactly cover the cost c.

VII. Sharing policies and incentives

In this section we analyse the inefficiency of simple sharing policies and their inability to optimally incentivize agents to contribute to the shared infrastructure.

For simplicity, we take the activity model of Section IV and assume that $\phi_i = 1$ for all i. Let the set of active agents at day t be S, where $\alpha(S) = \prod_{i \in S} \alpha_i \prod_{i \notin S} (1 - \alpha_i)$. Suppose $c(Q) = Q$ and that agents contribute daily q_1, q_2 (monetary or 'in kind') towards covering it, i.e. $Q = \sum_j q_j$. If all contending agents have the same concave utility function $u_i(x) = u(x)$, it would seem sensible to take $x_i(S) = Q/|S|$. But is this optimal? Or should the sharing policy depend on the

α_i and on the agents' contributions, q_i? One might expect that sharing resource amongst agents in proportion to their initial contributions provides better incentives and greater efficiency than sharing resource equally amongst agents. Next we analyze the performance of different simple policies for two agents, an equal sharing policy for n agents and subscription pricing in which all participants are charged the same fixed fee.

A. Sharing a resource between two agents.

Suppose $n = 2$. Let $x_i(S)$ be the share of resource given to agent i when the set of active agents is S. The average net benefit of agent 1 per period is

$$\alpha_1(1-\alpha_2)u(x_1(\{1\})) + \alpha_1\alpha_2 u(x_1(\{1,2\})) - q_1.$$

Suppose $u(x) = r - 1/x$, with $r = 10$, and $\alpha_1 = \alpha_2 = \alpha = 0.8$. If we take $x_i(\{i\}) = x_i(\{1,2\}) = q_i$ then we model agents acting alone, i.e. each building her own facility. Acting alone agent i maximizes $\alpha\,(r - 1/q_i) - q_i$. She obtains average net benefit of $\alpha 10 - 2\sqrt{\alpha} = 6.2112$, for $q_i = 0.8944$.

Now suppose agents share the resource. Since $\alpha_1 = \alpha_2 = \alpha$ we would expect that under any reasonable mechanism the agents should be incentivized to contribute equally and that resource should be shared equally when $S = \{1,2\}$. However, it matters what this mechanism is. We now look at such mechanisms.

Equal sharing. Consider an 'equal shares' policy of $x_i(\{i\}) = q_1 + q_2$ and $x_i(\{1,2\}) = \frac{1}{2}(q_1 + q_2)$. Agent i has net benefit of

$$nb_i(q_1,q_2) = \alpha\left(r - \frac{1-\alpha}{q_1+q_2} - \frac{\alpha}{\frac{1}{2}(q_1+q_2)}\right) - q_i.$$

The social optimum is achieved by choosing $q_1 = q_2 = q$ to maximize $nb_1(q_1,q_2) + nb_2(q_1,q_2)$. This is achieved by $q = \sqrt{\alpha(1+\alpha)} = 0.8485$. The net benefit per agent is 6.3029.

Suppose agents have full information regarding α, q_1 and q_2. Sharing resource with the equal shares policy, agent i maximizes $nb_i(q_1,q_2)$ with respect to q_i. There is equilibrium for any (q_1,q_2) such that $q_1 + q_2 = 1.2$. If we require $q_1 = q_2$ then the equilibrium is $q_1 = q_2 = 0.6$, and each agent has net benefit 6.2. This is less than the 6.2112 they obtain when acting alone. In fact, when $n = 2$, two identical agents will prefer to act alone for all $\alpha_1 = \alpha_2 > 7/9$.

The above issue worsens as the number of agents increases. If $n = 10$ then each agent contributes $q_i = 0.2561$ and the net benefit per agent is 5.1826. For $n \geq 98$ the equilibrium is driven to a point where agents no longer have positive net benefit. They will start deserting the system, even though, with a central planner, there would be benefit increasing in n.

We have made a surprising observation: two identical agents can obtain greater net benefit by acting on their own than by participating in a shared system in which their contributions are determined as the Nash equilibrium of a game. We have seen that the social welfare obtained by 'equal shares' can be less than stand alone for $\alpha > 7/9$. With $\alpha = 0.8$ the stand alone welfare is 6.2112 and the shared-infrastructure welfare is only 6.2. This is because the incentives are wrong and each agent tries to be a partial free-rider. How might we provide better incentives? One way is with proportional sharing.

Proportional sharing. Suppose we divide the resource between agents in proportion to their contributions. This gives $x_i(\{i\}) = q_1 + q_2$ and $x_i(\{1,2\}) = q_i$. The equilibrium is at $q_1 = q_2 = 0.8246$ and the social welfare is 6.30225, which is better than the stand alone welfare. This is just a bit less than the 6.30294 that a social planner could achieve.

Consider now a scheme that shares resource proportionally to sth powers of the contributions. That is,

$$x_i(\{i\}) = q_1 + q_2, \quad x_i(\{1,2\}) = \tfrac{q_1^s}{q_1^s + q_2^s}(q_1 + q_2).$$

Equal division is $s = 0$. Proportional division is $s = 1$. It turns out that the equilibrium point is increasing in s. For $s = 9/8 = 1.125$ the equilibrium is exactly the same as that of the social optimum. In fact, this works for any α when we take $s = \frac{1}{2}(1 + 1/\alpha)$. Note that this means taking $s \geq 1$.

Other schemes can also be good. For example, recall $q_1 = q_2 = q_0 = \sqrt{\alpha(1+\alpha)/2}$ achieves first-best welfare. Let

$$x_1(\{1\}) = q_1 + q_2 1_{\{q_1 \geq q_0\}}$$
$$x_2(\{2\}) = q_2 + q_1 1_{\{q_2 \geq q_0\}}, \quad x_i(\{1,2\}) = q_i.$$

That is, when agent 1 alone is active then she is allowed to use agent 2's contribution, but only if she contributes at least q_0. This scheme achieves the same social welfare as does a central planner. However, to compute q_0 we need to know the parameters α_1, α_2 (as when choosing $s = 1.25$ above).

B. Equal sharing provides wrong incentives.

The inadequacy of equal sharing is true more generally. Suppose that there are n agents, $\phi_i = 1$ for all i, and $\alpha_1 > \cdots > \alpha_n$. It turns out that the equal shares policy does not work well, because only agent 1 has any incentive to contribute resources. To see this, note that agent 1 wishes to maximize

$$nb_1(q) = \alpha_1\left[\alpha_2 Eu\left(\frac{\sum_i q_i}{M+2}\right) + \bar{\alpha}_2 Eu\left(\frac{\sum_i q_i}{M+1}\right)\right] - q_1$$

with respect to q_1, and agent 2 maximizes a similar expression $nb_2(q)$ with respect to q_2, where M is a random variable denoting the number of agents $3,\dots,n$ that are present. Since $\alpha_1(1-\alpha_2) > \alpha_2(1-\alpha_1)$ it follows that

$$\partial nb_1(q)/\partial q_1 = 0 \implies \partial nb_2(q)/\partial q_2 < 0.$$

So the only possible equilibrium is with $q_i = 0$, $i \geq 2$.

Now let M' be the number of the agents $2,\dots,n$ who are present. For an equilibrium to exist with $q_1 > 0$ and $q_i = 0$, $i \geq 2$, it would have to be that

$$\alpha_1 \partial E[u(q_1/(M'+1))]/\partial q_1 - 1 = 0$$

for some $q_1 > 0$. This can happen if and only if

$$\alpha_1 u'(0) E\left[1/(M'+1)\right] - 1 > 0.$$

Clearly, $E\left[1/(M'+1)\right] \to 0$ as $n \to \infty$. So if $u'(0) < \infty$ and n is large then no agent will wish to make any contribution.

C. Equal sharing with subscription pricing.

One possible scheme is to charge a flat subscription fee to any agent who wishes to participate. We purchase the greatest amount of resource that the collected fees allow, and in each epoch share it equally amongst any agents who are active. Such schemes are commonly used in practice due to their simplicity. Let us investigate how well such a scheme can do.

Suppose that $\phi_1 = \cdots = \phi_n = 1$, but α_i differ, and that a priori these are uniformly distributed on $[0, 1]$. If we set the fixed subscription fee to be q then there is a minimum α, say α_q, for which it is advantageous for a 'marginal' agent to participate. Suppose N is the number of the other $n-1$ agents who have their α_i greater than α_q. As the marginal agent's net benefit is 0,

$$0 = \alpha_q E_N \left[r - \frac{1 + \left(\frac{1+\alpha_q}{2}\right) N}{(N+1)q} \right] - q.$$

Using $N \sim B(n-1, 1-\alpha_q)$, routine calculation gives

$$0 = \alpha_q \left(r - [1 - \alpha_q^n + (1+\alpha_q)n]/(2nq) \right) - q,$$

and the expected net benefit of all the agents is

$$\tfrac{1}{2}(1-\alpha_q^2)n \left(r - \frac{1 - \alpha_q^n + (1+\alpha_q)n}{2nq} \right) - (1-\alpha_q)nq.$$

For $r = 10$ we find optimal q and α_q as in the table that follows. For comparison, the final column shows the first-best that could be obtained in the full information case. We can also calculate that under proportional sharing, as $n \to \infty$, agents of activity α are incentivized to contribute $\sqrt{0.6\alpha}$, and the average net benefit per agent is 3.967. Stand-alone it would be 3.667.

n	q	α_q	net benefit/agent	
			subscription	first-best
2	0.6367	0.0726	3.770	3.827
10	0.5418	0.0697	3.939	3.966
∞	0.5158	0.0575	3.987	4.000

Of course it would be even better to ask for a subscription fee that depends on α, which could then be policed. For example, this might be $\alpha\, q$. For n large it is optimal to take $q = 1$, there is no α_q, and the expected net benefit is $\approx 4n$, which is almost the same as using subscription $q = 0.5158$ for all agents. Other schemes might be investigated, such as sharing in proportion to q_i/α_i.

VIII. Building systems with many participants

We now address the formation of systems with large numbers of participants and show that optimal tariffs have a simple structure. Again our aim is to incentivize agents to report indirectly some private parameter by choosing the tariff that suits them most. As a function of his tariff choice, an agent is guaranteed a certain amount of service and the operator uses the payments to procure the infrastructure at the right size Q. In particular we consider tariffs of the form $\{(p(t), x(t)) : t \in [0, 1]\}$, parametrized by t, such that an agent who chooses tariff t, pays $p(t)$ and gets $x(t)$ whenever he is active. Each agent chooses the parameter t that offers him the best combination of cost and value. An equivalent non-parametric representation would be $x = x(p)$, a function of the payment.

Our analysis treats special cases of the general model in Section IV. Specifically, in Section VIII-A we deal with a new problem which we did not address before: we have the activity model, and the private information of an agent is his activity frequency α_i. In Section VIII-B the unknown parameter is service valuation ϕ_i.

A. Optimal incentives for declaring activity frequencies

We now consider the optimum designs for systems in which a large number of agents participate and which are of the activity model type introduced in Section IV, i.e. $\theta_{i,t}$ is 0 or 1, with probabilities $1 - \alpha_i$ and α_i respectively.

Suppose that $E\phi_i = 1$, which for simplicity we approximate as $\phi_i = 1$ for all i, and the values of $\alpha_1, \ldots, \alpha_n$, are unknown to the system designer. He would like to elicit these as part of an incentive compatible scheme that optimally sizes a system whose cost is covered by the payments of the agents. This model applies to the practical circumstance in which the central planner does not use accounting mechanisms to estimate, and thereby police the α_is. The aim is to structure the tariffs to incentivize agents to reveal truthfully their α_is.

Let us suppose that $c(Q) = Q$, $u(x) = \sqrt{x}$ and that a priori the α_i are distributed uniformly on $[0, 1]$. That is, there are approximately equal numbers of agents with each value of α in the range $[0, 1]$. The number of agents is very large, so we may suppose (by the law of large numbers) that we can meet demands from the common resource pool provided the total amount contributed through payments covers the cost of meeting average demand. The numbers we obtain in this section can be viewed as upper bounds on performance for a system with a small number of agents.

We would like to compare efficiency of the second-best policy with the full information case, but also with the case where agents use a different policy, the 'go-it-alone' policy, to self-provide their infrastructure and not share it with others.

The go-it-alone solution: If an agent with parameter α must go-it-alone then he will choose to build a facility of size x to maximize $\alpha u(x) - x$ and therefore take $x = \frac{1}{4}\alpha^2$. The average social welfare per agent is then

$$\int_0^1 \tfrac{1}{4}\alpha^2 \, d\alpha = \tfrac{1}{12} = 0.083\dot{3}.$$

The full information solution: Suppose a system designer having full information decides to provide an agent with parameter $\alpha = t$ with resource $x(t)$. The expected social welfare (per agent) is

$$\int_0^1 t\, u(x(t))\, dt - \int_0^1 t\, x(t)\, dt.$$

So the optimum is $x(t) = 1/4$ for all t, and the resulting social welfare per agent is $1/8 = 0.125$. It is somewhat surprising that a system designer will wish to allocate the same resource of $1/4$ to any agent on occasions he is present. This is because every time any agent is present he presents an opportunity to earn benefit $\sqrt{x}$.

The partial information solution using optimal tariffs: Now the designer of the system wishes to optimize the system by designing appropriate incentive compatible tariffs. Each agent chooses the tariff that is most beneficial to him. A tariff specifies the amount of resource an agent will receive each time he is active and the corresponding payment he must make initially in order to participate in such a system.

We consider the set of tariffs $(p(t), x(t))$ parametrized by t, the type of the customer (in this case α_i). According to these tariffs an agent who contributes $p(t)$ gets $x(t)$ whenever he is active, and $\{p(t), x(t) : t \in [0,1]\}$ is the set of possible choices. An agent's maximum net benefit is $f(\alpha)$, where

$$f(\alpha) = \max\left\{\max_s \left[\alpha u(x(s)) - p(s)\right], 0\right\}. \tag{26}$$

The maximum of linear functions of α is convex in α; this is how we know $f(\alpha)$ is convex. Similar to the arguments in Section V-B, for incentive compatibility we must have

$$\alpha u'(x(\alpha))x'(\alpha) - p'(\alpha) = 0.$$

So if an agent with parameter $\bar{\alpha}$ has net benefit 0, then incentive compatibility is equivalent with

$$p(\alpha) = \alpha u(x(\alpha)) - \int_{\bar{\alpha}}^{\alpha} u(x(s))\, ds$$

and

$$\int_{\bar{\alpha}}^{1} p(\alpha)\, d\alpha = \int_{\bar{\alpha}}^{1} (2\alpha - 1) u(x(\alpha))\, d\alpha. \tag{27}$$

The resource constraint is

$$\int_0^1 [\alpha x(\alpha) - p(\alpha)]\, d\alpha \le 0.$$

Our goal is to maximize the social welfare subject to incentive compatibility and cost coverage. Consider the net benefit from (26), the constraint (27), and substitute the resource constraint which holds with equality. Then we seek to maximize point-wise for each s a Lagrangian of

$$L = \int_{\bar{\alpha}}^{1} \Big[(s + \lambda(2s-1))u(x(s)) - (1+\lambda)s x(s)\Big]\, ds,$$

where $\bar{\alpha} = \lambda/(1+2\lambda)$ (the value of s such that $s + \lambda(2s-1) = 0$).

For $u(x) = \sqrt{x}$ the maximizing $x(s)$ is

$$x(s) = \left(\frac{2\lambda+1}{2(\lambda+1)} - \frac{\lambda}{2(\lambda+1)s}\right)^2.$$

This means that L is maximised to

$$\frac{1 - 2\,\lambda^2 \log\left(\frac{\lambda}{1+2\lambda}\right) - \lambda^2}{8(\lambda+1)}.$$

By minimizing this with respect to λ, we find $\lambda = 0.232206$. This gives for a solution in which for $\alpha \ge 0.158566 = \bar{\alpha}$,

$$p(t) = 0.173521 + 0.0942239 \log t$$
$$x(t) = (0.594224 - 0.0942239/t)^2$$

and $p(t) = x(t) = 0$ for $t < 0.158566$ $(= \lambda/(1+2\lambda))$.

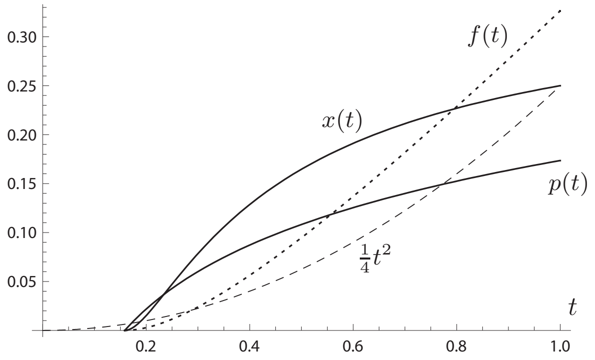

Fig. 2. The solid lines show $p(t)$ and $x(t)$ when $t > 0.2339$. The dotted line is net benefit $f(t) = tu(x(t)) - p(t)$ and the dashed line is $t^2/4$, the net benefit obtained by an agent acting alone.

Remarks.

1. The social welfare obtained is 0.116121 and this is just a bit less than the social welfare of 0.125 that could be obtained by a system designer having full information.

2. The optimal scheme is one in which agents with $\alpha \le \bar{\alpha} = 0.1586$ are prevented from participating. Intuitively, the reason we need to do this is so we can incentivize the agents with greater α to make more substantial contributions. Another way to think about this is that we prevent agents from free-riding by declaring small α, by preventing such small α from participating.

3. The black lines in Figure 2 show $p(t)$ and $x(t)$ (the amounts that agents will contribute and receive when declaring $\alpha = t$). Most agents receive more than they contribute. But agents with values of $\alpha \le 0.23389$ receive less than they contribute. However, if go-it-alone is not possible (because they cannot purchase and install resource for themselves or because there may be some additional fixed cost) then they will still take up this scheme, since their net benefit is positive.

4. The dashed blue line is $t^2/4$, which is the net benefit an agent could obtain if he were to go-it-alone, by taking $x(t) = p(t) = t^2/4$. The dotted red line is $f(t) = tu(x(t)) - p(t)$, the net benefit that an agent obtains in the shared system. This is convex, so there would be no benefit to an agent with parameter α masquerading as being two agents with parameters $\alpha/2$. Notice that the dotted red and dashed blue lines cross; an agent does better by going alone if $\alpha \le 0.2884$. It is easy to rework this analysis and obtain the optimal tariffs under the assumption that agents can go-it-alone if they find it more beneficial.

B. Optimal incentives for declaring service valuation

Now we look at infrastructures with a large number of participants and obtain a solution that is simple and intuitive.

We again consider the general model in Section IV, and specialize it for unknown ϕ_is, and $\theta_{i,t}$s which are truthfully reported because of policing and long-term participation. We analyze first the simple case in which agents are of the activity model type, i.e. $\theta_{i,t} \in \{0, 1\}$, with known activity frequencies. Then we generalize for arbitrary F_is. Thus we suppose that agent i is active on day t with probability α_i and when active and allocated resources x_i his benefit is $\phi_i u(x_i)$.

We state first a heuristic derivation of the large n result. For any n (not small) the optimal mechanism is like this: a system is built of size $Q(\phi)$. Agents are charged payments $p_1(\phi),\dots,p_n(\phi)$, and the sum of these covers the cost $c(Q(\phi))$. When agent i is contending for the resource amongst a group of active agents S he receives $x_i(\phi,S)$. Following the steps in the analysis presented in Section V-B, there is a $\lambda\geq 0$, such that for all S the optimal way to share resource Q amongst a set of active agents S with declarations ϕ is to maximize

$$\textstyle\sum_{i\in S}(\phi_i+\lambda g(\phi_i))u(x_i(\phi,S)), \tag{28}$$

over $\sum_i x_i(\phi,S)\leq Q(\phi)$. This follows from the fact that we are maximizing pointwise for each (ϕ,S), a Lagrangian of

$$L=E\Big[\textstyle\sum_{S,i\in S}\alpha(S)\big(\phi_i+\lambda g(\phi_i)\big)u(x_i)-(1+\lambda)c(Q)\Big]$$

over $Q=Q(\phi)$ and $x_i=x_i(\phi,S)$, subject to the constraint $\sum_{i\in S}x_i(\phi,S)\leq Q(\phi)$ for all S. The Lagrange multiplier λ is associated with the ex-ante constraint on cost coverage.

This has an interesting limit when n is large, and it allows payments to made in kind. Note that when agent i is active the rest of the system will be in its typical average state. So we can look for an approximate solution in which $x_i(\phi,S)$ is independent of S and we only need satisfy the constraint

$$\textstyle\sum_i\alpha_i x_i(\phi)\leq Q(\phi). \tag{29}$$

That is, we should choose the x_is so that the average sum of resource allocations does not exceed Q. Since $c(Q)=Q$, the Lagrangian for the problem reduces to

$$L=E\Big[\textstyle\sum_i\alpha_i(\phi_i+\lambda g(\phi_i))u(x_i)-(1+\lambda)\sum_i\alpha_i x_i\Big],$$

to be maximized with respect to $x_i\geq 0$. It turns out that as n becomes large, $\lambda\to 0$, the constraint (29) is satisfied, and the solution is

$$x_i(\phi)=x_i(\phi_i):=\arg\max_{x_i'}\{\phi_i u(x_i')-x_i'\}. \tag{30}$$

Moreover, each time an agent is active he is allocated the optimal amount independently of the other agents. But agent i pays only $\alpha_i x_i(\phi_i)$ and this exactly pays for his average resource usage. This is the form of first-best policy that one expects to obtain in the limit for large n, where Q is provisioned to serve the system in its typical (average) state. This suggests the simple form of optimal tariffs $(\alpha_i x,x), x\geq 0$, for each i. In this set of tariffs, agent i must choose a specific tariff, i.e. the best value of $x\geq 0$ given that by paying $\alpha_i x$ he obtains always, when he is active, x. Clearly, agents of large α_is should be policed so as not to use tariffs with smaller α_is.

It is interesting that the optimal contract chosen by agent i secures the same amount of resources from the shared resource pool as he would optimally choose to self-procure if no shared infrastructure was available and he was *always* active. But he needs only pay for his average usage, namely for $\alpha_i x_i$. By construction, this scheme is incentive compatible, i.e. he will choose the tariff parameterized by his actual value of ϕ. Note that $x_i(\phi_i)$ exceeds the size of the facility he would form if he were to stand-alone, which would be

$$x_i^0(\phi_i):=\arg\max_{x_i'}\{\phi_i\alpha_i u(x_i')-x_i'\}. \tag{31}$$

Thus an agent benefits from the existence of the other agents which are not always claiming resources; he uses the optimal amount when he is active but pays only when he uses it, since others pay for it when he is not.

In the general case the utility for agent i is $\phi_i\theta_{i,t}u(x_i)$. We make the assumption that $\theta_{i,t}$ is truthfully declared by policing and that it takes values from a finite set $\{\sigma_1,\dots,\sigma_M\}$, which is the same for all agents (for simplifying notation). Let $\alpha_{i,m}$ be the publicly known probabilities that agent i has $\theta_{i,t}=\sigma_m$. For instance, $\theta_{i,t}$ could take a value $\sigma_m\in\{1,2,3\}$ (perhaps corresponding to low, medium, and high). Let $x_i(\phi,\theta_t)$ be the allocation to agent i when the agents are in state $\theta_t=\theta_{1,t},\dots,\theta_{n,t}$.

It is easy to see that same analysis holds as before. Again, we assume when agent i is active with $\theta_{i,t}=\sigma_m$, the rest of the agents are in their typical average state. So it is reasonable to look again for an approximate solution in which $x_i(\phi,\theta_t)$ depends only on the value of $\theta_{i,t}$, i.e. is of the form $x_{i,m}=x_i(\phi,\sigma_m)$, and we only need to satisfy the constraint

$$\textstyle\sum_i\sum_m\alpha_{i,m}x_i(\phi,\sigma_m)\leq Q(\phi). \tag{32}$$

That is, we should choose the $x_{i,m}$s so that the average sum of resource allocations does not exceed Q. As before $x_{i,m}$ is the guaranteed amount agent i gets when his state $\theta_{i,t}=\sigma_m$. Again it turns out that as $n\to\infty$ we have a solution in which

$$x_{i,m}(\phi)=x_{i,m}(\phi_i):=\arg\max_{x_i'}\{\phi_i\sigma_m u(x_i')-x_i'\}. \tag{33}$$

This achieves the first-best optimum where each time an agent is in state m he is allocated the optimal amount $x_{i,m}(\phi_i)$ independently of the other agents and he pays only $\alpha_{i,m}x_{i,m}(\phi_i)$. Hence in total he pays for his average resource usage. This suggests the generalized form of optimal tariffs $(\alpha_{i,m}x,x)$, $x\geq 0$, for each i and m. In this set of tariffs, agent i must choose a specific tariff for each state m, i.e. the best value of x given that by paying $\alpha_{i,m}x$ he obtains always x when he is in state m. Clearly this tariff is possible because, as a result of policing, the value of m is truthfully declared.

Before stating a limiting result for the optimality of the above tariffs, lets see why it works. We work again using the easier notation of the activity model. Let $x_i=x_i(\phi_i)$, as defined above in (30). In practice, we need $\sum_i x_i(\phi,S)\leq Q(\phi)$ for all S. This is not possible if we try to take $x_i(\phi,S)=x_i$ for all S. However, we can modify things slightly by proposing an approximation of our previous tariff which is implementable for every n and in the limit becomes the tariff $(\alpha_i x,x)$. With agent i contributing q_i, we let $y_i=q_i/\alpha_i$ and redefine $x_i(\phi,S)=y_iQ/\sum_{j\in S}y_j$, where $Q=\sum_j q_j$. This is a proportional sharing of Q that takes into account the contributions of the agents and their frequency of use.

We illustrate the scheme with $u(x)=r-1/x$. Let $I_j\sim B(1,\alpha_j)$. Agent i has expected net benefit of

$$\begin{aligned}&\alpha_i\phi_i E\left[r-\big(y_i+\textstyle\sum_{j\neq i}I_jy_j\big)/(y_iQ)\right]-\alpha_i y_i\\&\qquad=\alpha_i\big(\phi_i(r-1/y_i)-y_i\big)-\alpha_i(1-\alpha_i)/Q.\end{aligned}$$

The term $\alpha_i(1-\alpha_i)/Q$ is small and varies little with y_i, and $\alpha_i(\phi_i(r-1/y_i)-y_i)$ is maximized by $y_i=x_i(\phi_i)$. So agent i is incentivized to contribute $\approx\alpha_i x_i$ and the total welfare, which is $O(n)$, will differ from its first-best value by just $O(1)$.

A limiting result for large n: Consider the activity model with n identical agents. Suppose that each agent is present with the same frequency α, and when present agent i has utility for resource $\phi_i u(x_i)$. Resource costs $c(Q) = Q$. Let us do everything ex-ante and for a single period. The aim is to maximize the expected net benefit (with the usual idea that we must satisfy C1–C3).

We know that the solution is one in which agent i will participate if $\phi_i + \lambda g(\phi_i) > 0$, and then, if the set of agents who turn out to present is J, the resource will be allocated to maximize $\sum_{i\in J}(\phi_i + \lambda g(\phi_i))u(x_i)$.

Let us define f_n^1 and f_n^2 as the maximal first- and second-best social welfares that can be obtained from n identical agents. We shall show that as $n \to \infty$ both f_n^1/n and f_n^2/n converge to the same social welfare per agent, as could be obtained if perfect multiplexing of resource allocations to agents were possible: meaning that if an agent were to make a contribution towards Q of αx and then he could receive precisely x whenever he is present. If this could be guaranteed then the agent with parameter ϕ_i should choose

$$x(\phi_i) = \arg\max_x \Big[\alpha\phi_i u(x) - \alpha x\Big] = \alpha \arg\max_x \Big[\phi_i u(x) - x\Big].$$

Let us define

$$z(\phi) = \max_x\{\phi u(x) - x\} = \phi u(x(\phi)) - x(\phi)$$

$$\bar{z} = E z(\phi).$$

Note that $f_n^2 \le f_n^1 \le n\alpha\bar{z}$.

Theorem 3: $f_n^2/n \to \alpha\bar{z}$.

Proof: We already have $f_n^2 \le n\alpha\bar{z}$. To establish an inequality in the opposite direction we need to find a mechanism that is implementable and which achieves a social welfare of almost $n\alpha\bar{z}$.

Let us suppose that ϕ_i has a distribution F over an interval $[a, b]$. Assuming $u(\cdot)$ is concave, then $x(\phi)$ is increasing in ϕ. Suppose that $0 < x(a) < x(b) < \infty$.

Fix some small $\epsilon > 0$. Suppose that it is possible to create a mechanism with the property that if an agent contributes $\alpha(1+\epsilon)x$ then he can be given an ex-ante guarantee that if he is present he will receive exactly resource amount x with probability $1-\epsilon$ and resource 0 with probability ϵ. Assuming this is so, agent i will choose to contribute an amount x_i which maximizes his ex-ante expected net benefit of

$$\alpha(1-\epsilon)\phi_i u(x_i) - \alpha(1+\epsilon)x_i,$$

and so he will take

$$x_i = x\left(\tfrac{1-\epsilon}{1+\epsilon}\,\phi_i\right).$$

Define

$$x_{\min} := x\left(\tfrac{1-\epsilon}{1+\epsilon}\,a\right) \quad \text{and} \quad x_{\max} := x\left(\tfrac{1-\epsilon}{1+\epsilon}\,b\right).$$

Assume $x_{\min} > 0$. Note that $x_{\min} \le x_i \le x_{\max}$.

We now show that for large n it is possible to fulfill the ex-ante guarantee to every agent. To see this, we observe that the probability we cannot provide resource of x_1 to agent 1 is no more than the probability that the total resource is insufficient to give resource of x_j to every agent j who is present. Letting I_i be a $B(1,\alpha)$ random variable, this is

$$\begin{aligned}
&P\left(x_1 + \sum_{i=2}^n I_i x_i > (1+\epsilon)\alpha x_1 + \sum_{i=2}^n (1+\epsilon)\alpha x_i\right)\\
&= P\left((1-\alpha)x_1 + \sum_{i=2}^n (I_i - \alpha)x_i > \epsilon\alpha x_1 + \epsilon\sum_{i=2}^n \alpha x_i\right)\\
&\le E_{x_2,\dots,x_n}\left[\frac{(1-\alpha)^2 x_1^2 + \sum_{i=2}^n \alpha(1-\alpha)x_i^2}{\Big(\epsilon\alpha\sum_i x_i\Big)^2}\right]\\
&\le \left(x_{\max}/x_{\min}\right)^2/\epsilon^2\alpha^2 n,
\end{aligned}$$

where for the first inequality we have used a Chebyshev inequality of the form $P(Y > \delta) \le E[Y^2]/\delta^2$, taking the expectation here over $I_2, \dots, I_n$.

Thus for n sufficiently large the right hand side above is less than ϵ, uniformly in x_1. Thus the ex-ante guarantee to agent 1 (and similarly other agents) can be fulfilled. The expected net benefit for agent i is then

$$\alpha(1+\epsilon)z\left(\tfrac{1-\epsilon}{1+\epsilon}\,\phi_i\right).$$

Now $z(\phi_i)$ is convex in ϕ_i. So by Taylor expansion in ϵ around 0, and using the fact that $z'(\phi_i) = u(x(\phi_i))$, we find that agent i has, for some $\epsilon_0 \in [0, \epsilon]$, expected net benefit of

$$\begin{aligned}
(1+\epsilon)z\left(\tfrac{1-\epsilon}{1+\epsilon}\,\phi_i\right) &= z(\phi_i) + [z(\phi_i) - 2\phi_i z'(\phi_i)]\epsilon\\
&\quad + 2\phi_i^2 z''(\phi_i)\epsilon_0^2\\
&\ge z(\phi_i) + [z(\phi_i) - 2\phi_i z'(\phi_i)]\epsilon\\
&= z(\phi_i) - [x(\phi_i) + \phi_i u(x(\phi_i))]\epsilon\\
&\ge z(\phi_i) - [x(b) + u(x(b))]\epsilon.
\end{aligned}$$

This shows that the expected net benefit that can be obtained from n agents by use of an optimal mechanism is at least $n\alpha\bar{z} - \alpha n[x(b) + u(x(b))]\epsilon$, for large n. Since ϵ is arbitrary, this completes the proof that $f_n^2/n \to \alpha\bar{z}$. ■

IX. Conclusions

We have investigated policies for running shared computing resource infrastructures. We have assumed that participants are strategic in disclosing private information about their actual resource needs and we have considered how best to share resources and take payments from the participants so as to maximize the overall efficiency of the system, while covering its costs. The chief lessons from this study are as follows.

1. A participant's decision about the quantity of resources that he will choose to contribute to the resource pool can be greatly affected by the resource allocation mechanism that he knows will be deployed when the system operates. Thus, a resource allocation policy may not be optimal if it only allocates resources with regard to the efficiency of the division of resources, while ignoring the effect this has on incentivizing agents to contribute towards covering cost. For example, if the resource will be shared equally amongst participants then an agent may choose to contribute nothing to the resource pool.

2. One way to incentivize potential participants to make significant contributions to the resource pool is to impose a rule that a participant will only be permitted to draw on the pool if he makes a minimum contribution to it at the point that it is formed. Another important rule is that an agent who contributes more resource will have greater priority for obtaining resource than an agent who has contributed less. Such rules will incentivize agents to make contributions that reflect their privately held beliefs about the benefits they expect to obtain. The result is a facility with an appropriately large quantity of resource, which is efficiently shared. Since contribution are made in kind there is no need for any internal money transfers.

3. We have seen that some optimal resource sharing mechanisms are parameterized by a Lagrange multiplier λ, or by the $\bar{\theta}$ or $\bar{\alpha}$ of a marginal participant. In practice, it would be useful if one could discover the right value of these parameters by some sort of on-line adaptation algorithm. We suggest that this could be an area for fruitful research.

4. In a facility that is already built and so has a fixed size (such as NRNs, National Grid Infrastructures), the running cost must be shared by charging the participants. In general, if the identities of the participants change over time, then our results for one-shot participation suggest that one should to operate a specialized mechanism in which participants receive resource shares according to their declared needs, while generating enough payments to cover running cost. In the scenario of long-term participation simpler policies exist, but at the added cost of implementing some accounting, such as policing of the α_i.

REFERENCES

[1] "e-Infrastructures Reflection Group," /www.e-irg.eu/images/stories/publ/finnishpresidency-recommendationsanddecisions.pdf.

[2] "Joint policy security group," proj-lcg-security.web.cern.ch/proj-lcg-security/security_policy.html.

[3] "European grid initiative," web.eu-egi.eu/,www.eu-egi.eu/blueprint.pdf.

[4] "The Distributed Grid Accounting System (DGAS)," www.to.infn.it/grid/accounting/main.html.

[5] "User's guides for the DGAS services," https://edms.cern.ch/document/571271.

[6] R. B. Myerson, "Optimal auction design," *Mathematics of Operations Research*, vol. 6, no. 1, 1981.

[7] ——, "Mechanism Design," in *The New Palgrave Dictionary of Economics Online*, 2nd ed., S. N. Durlauf and L. E. Blume, Eds. Palgrave Macmillian, 2009, www.dictionaryofeconomics.com.

[8] "gLite middleware," glite.cern.ch/,edms.cern.ch/file/722398/1.2/gLite-3-UserGuide.pdf.

[9] G. Borges, M. David, J. Gomes, J. Lopez, P. Rey, A. Simon, C. Fernandez, D. Kant, and K. M. Sephton, "Sun grid engine, a new scheduler for EGCE middleware," pubs.doc.ic.ac.uk/egee-sge-integration/egee-sge-integration.pdf.

[10] C. S. Yeo and R. Buyya, "A taxonomy of market-based resource management systems for utility-driven cluster computing," *Software Practice and Experience*, vol. 36, no. 13, pp. 1381–1419, 2006.

[11] R. Buyya, D. Abramson, and S. Venugopal, "The grid economy," *Proceedings of the IEEE*, vol. 93, no. 3, pp. 698–714, 2005.

[12] R. Buyya, D. Abramson, J. Giddy, and H. Stockinger, "Economic models for resource management and scheduling in grid computing," *Concurrency and Computation: Practice and Experience*, vol. 14, no. 13–15, pp. 1507–1542, 2002.

[13] C. Courcoubetis, M. Dramitinos, T. Rayna, S. Soursos, and G. D. Stamoulis, "Market mechanisms for trading grid resources," See [14], pages 58–72, 2008.

[14] M. W. Cripps and J. M. Swinkels, "Efficiency of large double auctions," *Econometrica*, vol. 74, no. 1, pp. 47–92, 2006.

[15] G. K. Dhananjay and S. Sunder, "What makes markets allocationally efficient?" *Quarterly Journal of Economics*, vol. 112, no. 2, pp. 603–630, 1997.

[16] R. Mason, C. Courcoubetis, and N. Miliou, "Grid economics and business models," in *Grid Economics and Business Models: 6th International Workshop, GECON 2009, Delft, The Netherlands, August 24, 2009*, R. B. J. Altmann and O. F. Rana, Eds., vol. 5745. Springer, 2009.

[17] "Distributed European Infrastructure for Supercomputing Applications," www.deisa.org.

[18] "Enabling grids for E-sciencE," www.eu-egee.org/.

[19] "Partnership for Advanced Computing in Europe," www.prace-project.eu/.

[20] "WLCG MoU documents," lcg.web.cern.ch/LCG/planning/planning.html#res.

[21] "The OneLab Project," www.onelab.eu.

[22] "The PlanetLab project," www.planet-lab.org.

[23] G. Christodoulou and E. Koutsoupias, "Mechanism design for scheduling," *Bulletin of European Association for Theoretical Computer Science*, vol. 97, pp. 40–59, 2009.

[24] N. Nisan, A. Ronen, and A. Ronen, "Algorithmic mechanism design," *Games and Economic Behavior*, vol. 35, pp. 166–196, 2001.

[25] J. Y. Yu and S. Mannor, "Efficiency of market-based resource allocation among many participants," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 6, pp. 1244–1259, 2007.

[26] C. Courcoubetis and R. R. Weber, "Incentives for large peer-to-peer systems," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 5, pp. 1034–1049, May 2006.

Costas Courcoubetis is Professor in Computer Science and heads the Network Economics and Services Group and the Theory, Economics and Systems Lab at the Athens University of Economics and Business. He received his PhD from the University of California, Berkeley and from 1982 until 1990 he was Member of the Technical Staff in the Mathematical Sciences Research Center at Bell Laboratories. His research interests include economics of communication networks, resource allocation and optimization, peer-to-peer computing, and regulation policy. He is a co-author, with Richard Weber, of the book *Pricing Communication Networks: Economics, Technology and Modeling* (Wiley, 2003). He has been frequently a consultant with Bell Laboratories and the Greek regulation authorities, and has participated in many EU research projects related to network economics such as Ca$hman, M3i, MMAPPS, Gridecon and Etics

Richard Weber is Churchill Professor of Mathematics for Operational Research in the Mathematics Department of the University of Cambridge. His research interests include the economics of networks and pricing, bin packing, Gittins index, queueing control, rendezvous search, and scheduling. He is a coauthor (with Gittins and Glazebrook) of *Multi-armed Bandit Allocation Indices* (Wiley 2011). He has participated in EU research projects related to network economics such as Ca$hman, M3i, MMAPPS and Gridecon.