MANAGEMENT SCIENCE
Vol. 51, No. 7, July 2005, pp. 1141–1150
ISSN 0025-1909 | EISSN 1526-5501 | 05 | 5107 | 1141

DOI 10.1287/mnsc.1050.0419
© 2005 INFORMS

Pricing and Allocation for Quality-Differentiated Online Services

Ravi Bapna, Paulo Goes
Department of Operations and Information Management, School of Business, University of Connecticut, Storrs, Connecticut 06269 {ravi.bapna@business.uconn.edu, paulo.goes@business.uconn.edu}

Alok Gupta
Information and Decision Sciences Department, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455, agupta@csom.umn.edu

We explore the problem of pricing and allocation of unique, one-time digital products in the form of data streams. We look at the short-term problem where the firm has a capacitated shared resource and multiple products or service levels. We formulate the allocatively efficient Generalized Vickrey Auction (GVA) for our setting and point out the computational challenges in determining the individual discriminatory transfer payments. We propose an alternative uniform-price, computationally efficient, revenue-maximizing knapsack formulation called the Multiple Vickrey Auction (MVA). While not incentive compatible, the MVA mechanism achieves bounded posterior regret and can be solved in real time. It has the added benefit of realizing imputed commodity prices for the various services, a feature lacking in the discriminatory GVA approach. For service providers that are concerned about the incentive compatibility but want imputed service prices, we suggest a maximal MVA (mMVA) uniform-pricing scheme that trades off revenue maximization for allocative efficiency. For sake of completeness we discuss the properties of a first-price pay-your-bid scheme. While NP-hard and not incentive compatible, this formulation has the perceived benefit of cognitive simplicity on the parts of sellers and bidders.

Key words: quality-of-service-oriented digital services; auctions to reveal valuations of one-time services; nonstandard knapsack formulation
History: Accepted by Rajiv D. Banker, information systems; received April 4, 2003. This paper was with the authors 5 months and 1 week for 3 revisions.

1. Introduction and Background

We consider a joint preference-elicitation, pricing, and capacity allocation problem faced by a firm offering unique, one-time digital products in the form of data streams. Such products include, for example, real-time webcasts of events, video on demand, or videoconferencing sessions. The firm operates servers of limited capacity that are able to generate digital streams at different quality levels. The capacity is usually associated with a discrete number of connections established between the server and the customers' clients. Servers are connected to the clients through a network. Typical connections can take the form of circuits in dedicated networks, virtual circuits in asynchronous transfer mode (ATM) like implementations, or persistent TCP connections to the Internet. The data stream products are consumed in real time with consumers having choices of different guaranteed quality levels under which they purchase them. Our research challenge is to design a preference-elicitation scheme and couple that with a pricing and allocation scheme. Our decision variables determine how many subscribers to accept in a given service class and at what price, given a fixed number of service classes and a fixed capacity.

Note that we consider the servers' location to be on the edge of the network. It is important to note that in this paper we *do not* explicitly consider problems such as the data transmission delays over the Internet. Our analysis and proposed market mechanism focuses on the *server side*. In the current Internet implementation, network latency is jointly minimized by locating such servers at the edge of the network (for instance, through akamai.com), as well as the adoption of broadband connection technologies. Irrespective of where the server is located, the content providers still need to be able to structure markets to generate revenue, recover technology costs, and ultimately maximize profits, or in certain cases, allocative efficiency. We believe that the creation of mercantile processes for such real-time unique services is one of the largest areas of Internet expansion that has been overlooked.

We contribute by providing a menu of preference-elicitation, pricing, and allocation schemes, along with their trade-offs. We begin by formulating the

Generalized Vickrey Auction (GVA) for our problem setting, and subsequently propose a uniform-price Multiple Vickrey Auction (MVA) alternative that is computationally efficient. While the GVA is known to be allocatively efficient and incentive compatible, it presents computational challenges in determining individual transfer payments. The MVA has the added benefit of realizing imputed commodity prices for the various services classes, a feature lacking in the discriminatory GVA approach. For service providers that are concerned about the incentive compatibility but want imputed service prices, we suggest a maximal MVA (mMVA) uniform-pricing scheme that trades off revenue maximization for allocative efficiency. For sake of completeness we discuss the properties of a first-price pay-your-bid scheme. While NP-hard and not incentive compatible, this formulation has the perceived benefit of cognitive simplicity on the parts of sellers and bidders. We believe that by understanding the theoretical, computational, and economic property trade-offs, service providers can make informed decisions regarding which mechanism to use in the context of their markets.

1.1. The Capacitated Server

We present three motivating examples of capacitated-server electronic marketplaces. In such markets, the seller has to decide what data streams to offer, at what price, to which consumers, and at what quality. The three exemplifying markets are Internet webcasts of one-time events, video on demand, and (more generally) the pricing and allocation of ATM traffic. Interactive webcasts of concerts, high-profile interviews, and sporting events such as international soccer and cricket matches[1] are streamed from multimedia servers, typically located at the edge of a network. In working with streaming media, a general rule of thumb is to provide 12 KB of random-access memory for every kilobit per second of streaming (RealSystem 2001). Similarly, video-on-demand providers constantly deal with the issue of optimally managing the server capacity, or the "content shelf-space" as it is referred to in the industry. These firms serve as intermediaries between content providers, such as ESPN and Disney, and consumers in specialized high-bandwidth digital cable markets. The video-on-demand server's capacity is the critical link in maintaining the desirable service quality level with broadband broadcast technologies.[2] Our approach is also applicable in a more general setting that involves pricing and allocation decisions for an operator of an ATM network. In ATM networks, quality of service (QoS) refers to specific traffic-handling parameters that are adhered to for a given circuit. Managing such networks efficiently is always a challenge because of the enormous number of choices available for setting the various operational parameters. The findings of this paper can be used to derive an "optimal" bandwidth allocation plan together with a pricing system for the real-time delivery of ATM services, such as videoconferencing and dynamic content.

In our models we require capacity to be reserved for each consumer for the duration of the transmission of the digital streams. We recognize that solving the server-side resource allocation problem is a necessary, but insufficient, condition to make these markets feasible. Joint consideration with network latency would complete the picture. However, as we demonstrate, this problem in itself has significant intractability that has not been addressed by the literature. It provides the foundation to build the structure.

In the current scenario, given that servers can support only a finite number of connections simultaneously, services are either provided free with no QoS guarantees or, more recently, using posted-price mechanisms.[3] These markets are in their infancy, and it is safe to assume that consumers will be reluctant to pay for services that are often delivered extremely poorly with frequent breakdowns and overloads. The problem arises from the difficulty of assessing consumers' willingness to pay and allocating appropriate aggregate capacity at servers to deliver desired/required QoS. Our research provides the first steps by integrating a fair, uniform-price (for a given product), auction mechanism with a combinatorial optimization–based capacity allocation that maximizes sellers' revenue and facilitates QoS guarantees by reserving server capacity for each consumer.

We assume a monopolistic seller. Note that this is not a very restrictive assumption due to the unique, one-time nature of products being considered. Transmission rights have to be obtained from the primary holder of the media rights, thus, the provider typically does not face market competition for the delivery of such services.

We next motivate the need for an auction-based preference-elicitation scheme.

1.2. Auction-Based Preference Elicitation

There are three primary types of economic resource allocation mechanisms: (i) capacity allocation mechanisms, (ii) posted-price mechanisms, and (iii) auctions and negotiations. Capacity allocation mechanisms

[1] www.willow.tv offers different quality levels by offering different transmission speeds out of the servers.

[2] www.divatv.com has operationalized such a delivery mechanism by providing 3.4 Mbps MPEG2 streams that can be delivered with less than 0.8-second latency by each of their servers.

[3] Examples include Cricket matches on www.willow.tv and University of Connecticut women's basketball matches on www.cptv.com.

usually are the most efficient mechanisms if the type of individual customers, and thus, their needs, can be identified by the controlling entity. In general, posted-price mechanisms can be considered as the mathematical dual of capacity allocation mechanisms (Greenwood and McAfee 1991). Under this mechanism, the general distribution of customer types is known, but individual customer type is not identifiable. In other words, the demand curve is known. Even though some research has been devoted to developing posted-price mechanisms that can dynamically compute the prices based on changing demand (see, for example, Gupta et al. 1997), these mechanisms are more effective when a relatively long-term demand trend is available. In this paper, we use mechanisms based on auctions because we are considering the allocation of resources for unique one-time products and services, the demand for which may not be assessable in advance. Given the context, it is well known that a posted-pricing mechanism may not be optimal (Wang 1993). It is worth mentioning that in the absence of spatial, temporal, and geographic constraints, online auctions are being increasingly adopted on the Internet. They are also attracting renewed interest in the vast auction literature (see Lucking-Reiley 2000 for a review of what is being auctioned, Bapna et al. 2003 for an analysis of multiunit business-to-consumer online auctions). We formulate and contrast four auction-based preference-elicitation and pricing and allocation schemes in §2. The primary contribution of this study is to posit an auction-based market structure for eliciting preferences for information goods, followed by a knapsack formulation to determine allocation and pricing. To the best of our knowledge, this approach has not been studied in the literature.

2. Model Formulation and Structural Analysis

2.1. Assumptions

We assume that customers have a value for service that is unknown to the provider. Further, we assume that the provider will use some price-setting mechanism and some customers may be excluded from receiving the service based on the prices they are willing to pay.

Let there be $i = 1, \ldots, I$ consumers in the market for $j = 1, \ldots, J$ different services offered by a provider. Let V_{ij} denote customer i's valuation for service j. We assume that consumer's valuations V_{ij} are independently distributed.

Let C represent the total bandwidth, or capacity, and let C_j be the capacity consumed by service j. Finally, let x_{ij} represent the decision variables where $x_{ij} = 1$ if customer i receives service j, and $x_{ij} = 0$ otherwise.

We assume that given a price-setting mechanism and the negligible marginal cost of offering the digital services under consideration, a provider's objective is to maximize either revenue or allocative efficiency. We begin with the well-known GVA approach. The main idea here is to achieve allocative efficiency and to reward bidders for truth telling by giving them discounts equal to what they bring to the table—namely their marginal product.

2.2. The Generalized Vickrey Auction Approach—A Primal-Dual Model

Along the lines of the model proposed by Bikhchandani and Mamer (1997), we begin by exploring the structure of a potential Walrasian equilibrium of the market through a primal-dual formulation. We say "potential" because it is not certain whether such an equilibrium exists. The primal's objective is to maximize allocative efficiency—that is, allocate the goods to the agents who value them the most—while the dual is useful in determining incentive-compatible prices. We introduce a variable y_l for every feasible supply vector $l = (S_1^l, \ldots, S_J^l)$, such that $\sum_{j\in J} S_j^l = C$. Then the economy can be modeled as

$$\text{Maximize} \quad \sum_i \sum_j x_{ij} v_{ij}, \tag{1}$$

$$\text{subject to} \quad \sum_j x_{ij} \le 1 \quad \forall\, i = 1, \ldots, I, \tag{2}$$

$$\sum_i x_{ij} c_j - \sum_l y_l s_j^l \le 0 \quad \forall\, j = 1, \ldots, J, \tag{3}$$

$$\sum_l y_l \le 1, \tag{4}$$

$$x_{ij}, y_l = \{0, 1\}. \tag{5}$$

The primal's objective function maximizes the allocative efficiency. Constraints (2) and (3) represent the assignment constraint and the capacity constraint given a chosen allocation, respectively. The corresponding dual of the linear programming (LP) relaxation of (1)–(5) is

$$\text{Minimize} \quad \sum_i \pi_i + \pi_0, \tag{6}$$

$$\text{subject to} \quad \pi_i + p_j c_j \ge v_{ij} \quad \forall\, i, j, \tag{7}$$

$$\pi_o - \sum_j p_j s_j^l \ge 0 \quad \forall\, l, \tag{8}$$

$$\pi_o, \pi_i, p_j \ge 0. \tag{9}$$

Observe that the dual variable π_i corresponds to the bidders surplus, and p_j corresponds to the unit price for the service class j. The variable π_o corresponds to the seller's revenue and the resultant discriminatory prices for bidder i in class j, computed as $(v_{ij} - \pi_i)$, are nonlinear. For similar economies, and the special case where the LP relaxation is integral, Bikhchandani and Ostroy (2002) develop a primal-dual algorithm that resembles an ascending-price auction. In further special cases, these are truth revealing. The pricing rule for such an allocation scheme ensures that each winning bidder pays an amount equal to the sum of the losing bids that would have won had she not bid. Each bidder has the incentive to reveal their true valuations, driven by the discount in the form of the marginal contribution that is given to them.

2.2.1. Generalized Vickrey Auction Analysis. It is important to note that the model of our problem is a more general case because of the capacity constraints. If indeed a relaxed LP model corresponding to (1)–(5) results in integer solutions, as it does in some instances of the underlying package assignment problem (Bikhchandani and Ostroy 2002), then we have a computationally efficient Vickrey-Clarke-Groves (VCG) mechanism (Vickrey 1961, Clarke 1971, Groves 1973).

Regrettably, such a nicety is not forthcoming in this problem's context because of the presence of constraint set (3), the set of capacity constraints. Intuitively, the problem context is slightly outside the traditional combinatorial auction (package assignment, or in the case of a single seller, set packing) literature, where the number of distinct objects is known a priori. In the context of this study case, the number of bidders to accept in each class, i.e., the allocation, is a priori unknown (a decision to be optimized). Thus, while the primal-dual formulation is elegant in theory, it has significant operational and computational difficulties. Fractional solutions of the relaxed primal are certain and the very existence of Walrasian equilibrium is challenged in the problem formulation (1)–(9). Computing GVA prices would involve solving $I+1$ integer programs (Bikhchandani and Ostroy 2002), and would be realistic only in the context of small problems or in the absence of time sensitivity.

Another significant concern, also raised by de Vries and Vohra (2003) in their excellent review of combinatorial auctions, is that the formulation itself requires the enumeration of all feasible solutions. While the formulation (1)–(9) is useful for deriving economic insights and may provide a framework for designing ascending-price auctions in the context of capacity constraints, what this auction should look like is an open research question.

Next, we suggest an alternative formulation that is based on a sealed-bid uniform-price auction with a payment that is similar to the Vickrey payment.

2.3. The Multiple Vickrey Auction Approach

If a service provider has limited computational resources and/or has the need to for real-time decision-making capability, we propose an alternative knapsack formulation in the context of Vickrey-based prices. The "knapsack" is a capacitated server to be filled up with a discrete mix of services, and our problem has to have an additional nonlinear constraint for auction-based price discovery. The basic idea here is that within a service class, the resulting allocation must be fair. We operationalize fairness based on Foley's (1967) characterization of an envy-free allocation. A fair allocation scheme is defined as one in which there is no instance wherein, for the purposes of revenue maximization, a consumer with a lower revealed valuation for a given service can be allocated while there exists another, with a higher revealed valuation, for the same service who is denied. We call such an auction scheme an MVA.

The new constraint imposes the structure of a uniform-price Vickrey (1961) auction. In such auctions, if there are m units of good for sale (i.e., supply is a priori known), then the m highest bids each win, and the $(m+1)$st bid becomes the price paid for each unit won. A seminal result is that, in the absence of multiunit demand, which is the case in our market, such a mechanism induces truth telling from the bidders. In case bidders have multiunit demand, then Ausubel (2004) has shown that the incentive compatibility result has to be generalized to the GVA. Using Vickrey's uniform-pricing rule (as is) could lead to demand reduction and consequent inefficient allocation of resources. Yet, while theory suggests demand reduction, a recent empirical study found no evidence of it in a realistic field experimental setting. List and Lucking-Reiley (2000), contrast the Vickrey uniform-pricing approach with the GVA and find that as the number of bidders increases, both the mechanisms yield statistically equivalent expected revenue. We would expect the effect of large numbers to be present in the context of the Internet markets we are considering.

The first caveat for a service provider considering a fair approach—uniform price in a service class—is that we cannot guarantee incentive compatibility given an unknown number of units to sell in a given class. However, what we gain is an alternative approach of attacking the problem from a knapsack perspective and significant computational advantages that go along with it. If each customer can only bid for one product, and the number of units being sold is known, the special pricing structure thus obtained has an incentive compatibility property for customers as noted by Vickrey (1961). Because of this desirable property, the MVA has been suggested as a price-setting mechanism for online environments by several researchers (for example, see MacKie-Mason and Varian 1995, Lazar and Semret 1999).

For the remainder of this subsection, we assume that bidders accepted within a service class pay a uniform, to-be-determined price and prices can differ between service classes, reflecting a market view of the underlying commodity (bandwidth) bundles that make up service classes.

Let a surplus-maximizing consumer i bid $B_{ij} \leq V_{ij}$, where B_{ij} is customer i's declared bid for service j. The revenue-maximizing MVA knapsack formulation with Vickery uniform pricing is as follows:

$$\text{Maximize } z = \sum_i \sum_j x_{ij} p_j, \tag{10}$$

subject to

$$p_j x_{ij} \leq B_{i+1\,j} \quad \forall\, i = 1, \ldots, I, \forall\, j = 1, \ldots, J, \tag{11}$$

$$\sum_j x_{ij} \leq 1 \quad \forall\, i = 1, \ldots, I, \tag{12}$$

$$\sum_i \sum_j x_{ij} C_j \leq C, \tag{13}$$

$$x_{ij} = \{0, 1\}, \tag{14}$$

$$p_j \geq 0. \tag{15}$$

2.3.1. Multiple Vickrey Auction Analysis. Atypically, in the above formulation, there are two unknown quantities, x_{ij} and p_j, in the revenue-maximizing objective function represented by Equation (10). It is important to define the price-setting term $B_{i+1,\,j}$ for each class. The price is equal to the bid that is just after the last accepted consumer in a given class—hence the term $B_{i+1,\,j}$. In the case that the last bid in a class is accepted, we assume it to be the price-setting bid. Note that this problem (10)–(15) has the elements of the general 0–1 multiple-products knapsack problem with the additional constraints (11) and (12) that represents the Vickrey pricing constraint and the assignment constraint, respectively. Because both x_{ij} and p_j are decision variables being maximized, Equation (11) represents a nonlinear constraint ensuring that Vickery prices are set for each service class. Equation (12) ensures that any consumer can win in at most one service class. Equation (13) is a typical knapsack capacity constraint.

Both knapsack problems and quadratic assignment problems (Ahuja et al. 2000) on their own are known to be NP-hard. Yet, a wide variety of computationally efficient solution techniques, based on variants of the *greedy algorithm*, which is based on choosing the elements in nonincreasing value-to-weight (bids-to-capacity capacity) ratio (Martello and Toth 1989) have been developed to solve knapsack problems. For this paper we assume that such an algorithm exists.

Interestingly, our problem's structure has a further nicety that lends to pseudo-polynomial complete enumeration. The Vickrey uniform-pricing introduces dependencies between the bids of different individuals in the form of positive network externalities. Each new lower-value customer who is accepted into a particular service class lowers the price for all previously accepted customers of that class. Hence, the consumer's surplus $(V_{ij} - p_j)$ becomes a nondecreasing quantity. We leave it to the reader to verify that under such a structure the optimal solution can be obtained in $O(I^J)$. Thus, from a practical perspective, the MVA is computationally efficient.

Recent work on multiunit online auctions has established that both discriminatory English auctions, as well as the MVA, generate comparable revenues on average (Bapna et al. 2000). The MVA implementation proposed here has a significant advantage in that it does not discriminate among the winners with respect to the price they pay. This leads to imputed commodity prices for each of the service classes offered, a desirable feature in any market. We expect this mechanism to be more attractive to bidders and, as Vickrey (1961) noted, it reduces the probability that a bidder's own bid will affect her own price, making the revelation of values more likely. The simplicity associated with the uniform-pricing mechanism is likely to have an intuitive appeal to decision makers in online markets, but it's hard to quantify that effect. In the next subsection we discuss the equilibrium properties of the MVA mechanism.

2.3.2. Multiple Vickrey Auction Mechanism's Equilibrium Properties. It is useful to consider if, after the allocation, all bids and allocations were revealed to everyone, would any bidder have an incentive to move, assuming the others stayed constant? This consideration is based on the market-microstructure literature's focus on having posterior regret-free allocations (O'Hara 1995, p. 60). This "regret-free" price is consistent with the property found in rational expectations models incorporating the information revealed by the allocation itself. It is clear that in the presence of a variable number of item units (service class in our context), ex post manipulability is possible when the $(m+1)$st uniform-pricing rule is used. Bidders can ex post determine lower bids that would have, given a larger number of units, yielded a maximal solution to a revenue-maximizing provider, and hence resulted in an allocation for them at a lower price. It is also obvious that this issue is only present at the margin, and hence we expect that its impact is likely to be minimal. We formally derive the conditions necessary for realizing regret-free allocations. For expositional ease, but without any loss of generality, we assume a single service class.

Lemma 1. *In any optimal allocation using MVA the following must be true*

$$xB_{x+1} \leq NB_{N+1} \quad \forall\, x, \tag{16}$$

where N is the number of bidders receiving the allocation using MVA.

Proof. By construction, because allocating to N customers maximizes the revenue.

Theorem 1. *The MVA mechanism is posterior regret free if*

(i) *all the available capacity is used;*

(ii) *if there exists capacity for $k-1$ additional bidders, then the sufficient condition for the mechanism to be regret free is*

$$(N-1)/(N+k-1)B_{N+1} > B_{N+k}, \quad \forall k. \tag{17}$$

Proof. (i) If there is no capacity left, then the final allocation will be to N consumers at most. To win, any bidder has to bid B_{N+1} (the original price) or more. Therefore, no winning bidder can change their bid and affect their own price.

(ii) Note that a winning bidder $l < N$ can benefit only if he is able to have an allocation at a new lower price. Bidder l cannot affect the price by moving to any position $< N+1$ because the price still will be B_{N+1}.

If the bidder bids in any position from $N+m$, where $m \in [1, N+k-1]$, then from Lemma 1, the following has to be satisfied

$$(N-1)B_{N+1} \leq (N+m)B_{N+m+1}. \tag{18}$$

Note that the left term is the revenue from the original price now applied to $N-1$ bidders since the first bidder has moved down to $N+m$ position, whereas the term on the right is the revenue from $N+m$ being allocated with the new bid of bidder l. The mechanism will be posterior regret free if the condition above is not satisfied for any value of m. Thus, if there exists capacity for $k-1$ additional bidders, then the sufficient condition for the mechanism to be regret free is

$$(N-1)/(N+k-1)B_{N+1} > B_{N+k}, \quad \forall k. \quad \text{Q.E.D.}$$

Another point worth noting is that the condition (ii) of Theorem 1 reflects an interesting trade-off between the steepness of the valuation distribution at the margin, in particular between B_{N+1} and B_{N+k}, and on the tightness of the capacity constraint. Intuitively, a smaller spread in the valuation distribution is more likely to yield results where the allocation will be to full capacity and hence the mechanism is more likely to be regret free. However, as the spread increases, the dispersion in bids will be greater and the mechanism will again start converging toward being regret free.

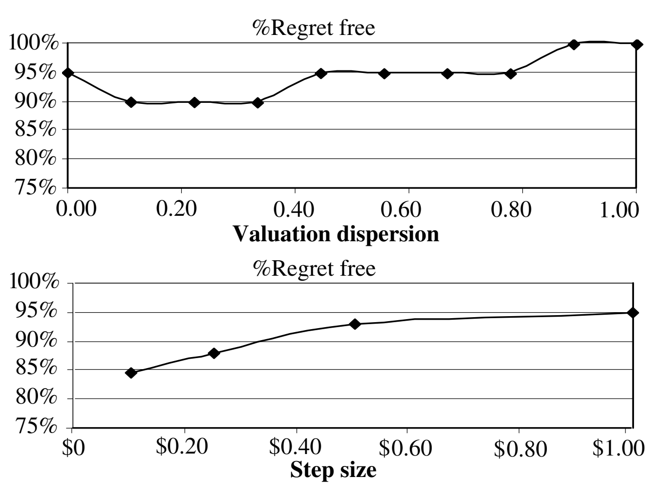
It is also interesting to note that the optimal choice of the cutoff N implies that B_{N+1} is *strictly greater* than B_{N+m}, where $m \in [1, N+k-1]$. This insight can be used to increase the likelihood of achieving posterior regret-free allocations and prices by setting a minimum step size that would correspond to the steepness of the valuation distribution at the margin. Formally:

Proposition 1. *If the optimal MVA allocation results in left-over capacity for $k-1$ more allocations, then $B_{N+1} > B_{N+m}$, $\forall m \in [1, N+k-1]$.*

Proof. The proof is by contradiction. If this were not true, the B_{N+m-1} bidder would have been included in the original allocation, because if for example $B_{N+1} = B_{N+2}$, then the $(N+1)$st bidder would have been included. This would happen because if B_{N+1} had a positive marginal contribution, then, by equality, so would B_{N+2}. Q.E.D.

Proposition 1 can be used directly by auctioneers to set a minimum step size in the bidding rules such that the likelihood of obtaining the regret-free condition of Theorem 1 is increased, because the relationship between B_{N+1} and B_{N+k} can be manipulated by the step size. An increase in step size will increase the likelihood of regret-free allocation because the difference between B_{N+1} and B_{N+k} will be larger, hence resulting in increased likelihood of satisfying Equation (17). To demonstrate this effect, we designed a simple, one-class, enumerative experiment to isolate the trade-off between the valuation distribution spread and the regret-freeness of the MVA. For the top graph of Figure 1, bidder valuations were drawn for all instances from a uniform distribution centered on a mean of \$10. However, we increased the dispersion by changing the upper and lower bound of the distribution by \$0.10 up to a maximum of \$1. The step size was fixed at \$0.50. The U-shaped top graph in Figure 1 starts off at a high valu e of regret-freeness for the low dispersion case, reflecting that the capacity

Figure 1 **Regret-Freeness as a Function of Valuation Dispersion and Step Size**



constraint is binding. The percentage regret-freeness drops initially, and then picks up as the increased dispersion dominates the effect of the capacity constraint because the difference between B_{N+1} and B_{N+k} is higher due to dispersion. The graph on the bottom of Figure 1 indicates that percentage regret-freeness increases with the step size, going from 85% for a \$0.10 increment to 95% for a \$1 increment. For both graphs, each point is an average of 30 replications.

Next, we consider a mechanism wherein a provider chooses to allocate all the available capacity, irrespective of revenue, while preserving fairness in allocation.

2.3.3. Maximal MVA Enforces Incentive Compatibility at Cost of Revenue. For service providers that desire an incentive-compatible uniform-pricing approach and associated commodity prices for each of the services the MVA formulation's (10)–(15) objective function (10) can be tweaked to maximize the number of jobs accepted instead of maximizing revenue. Thus the service provider optimally fills up the knapsack till capacity, enforces fairness, and charges all accepted bidders the price of the first rejected bid in their service class. We call such a scheme an mMVA allocation.

Proposition 2. *The mMVA mechanism is efficient and incentive compatible.*

Proof. Consider a non-revenue-maximizing uniform-price auction, where bidders are accepted until the capacity is filled. Efficiency follows by definition as goods are allocated to bidders that value them the most, and no bidder is left out if any capacity exists. Incentive-compatibility proof is along the lines of Vickrey (1961). Given a buyer's valuation, we only have to consider (1) the case of an accepted buyer's incentive to reduce her bid and (2) the case of a rejected buyer's incentive to increase her bid.

Given a maximal allocation, in Case 1, the buyer cannot benefit by decreasing the bid as it has no effect on the price for that class. Moreover, he can potentially lose the bid if the new bid is less than the price-setting bid. For Case 2, the buyer would end up paying more than her valuation to win, which would violate individual rationality. Q.E.D.

The mMVA auction scheme has the benefits of computational and allocative efficiency and incentive compatibility. From a commercial service provider's perspective, its only downside is that it is not revenue maximizing. Thus, if a service provider's objective is to maximize social welfare (say, a nonprofit organization) and have imputed prices, we would suggest the mMVA scheme. We are not aware of any comparative studies that compare the expected revenue of the GVA with the mMVA. This promises to be fertile ground for future lab and field experimentation.

In the next subsection we examine the properties of the mechanisms on the provider's revenue. Such analysis has been largely ignored in the literature.

2.3.4. Provider-Side MVA Mechanism Analysis. We begin by considering the structure of the provider's revenue curve.

Observation 1. The revenue curve generated by the MVA is nonmonotonic.

Proof. Here it is sufficient to show that there exists an instance for which a content provider may actually realize less revenue by selling $Q+1$ items than selling Q items. Consider only one level of service requiring one unit of capacity, a total of $N=4$ units to be sold and a final sorted bid sequence of B, $B/2$, $B/5$, $B/11$, $B/11$. The provider receives a total revenue of $B/2$ if $Q=1$, $2B/5$ for $Q=2$, $3B/11$ for $Q=3$, and so forth. Clearly, in this case, the provider's revenue is decreasing with increasing number of units sold. Therefore, MVA does not necessarily generate a nondecreasing revenue curve. Q.E.D.

Observation 1 indicates that a revenue-maximizing content provider using the MVA will not commit to selling the maximum possible units of service. Instead, the provider may choose not to utilize all its available capacity. In our model where several services share a common capacity, the provider does not have to commit to a specific number of acceptances for any service. Thus, any formulation of such a mechanism must jointly consider how many bidders to accept in each service class and at what prices.

It is obvious, given the nonmonotonicity of the revenue curve, that a simple scan procedure that stops at the first negative swing in the marginal revenue is not appropriate to determine the optimal number of items to be sold. A single scan of the entire total revenue curve will result in an optimal decision, when there is more than one service, but such an approach will result in exponential combinations to be evaluated. However, the following key observation will help create an efficient method for searching for an optimal solution with multiple services.

Observation 2. If providing service to $N' \leq N$ consumers maximizes a provider's revenue for a given service under the MVA, it does not imply that the provider will consider it optimal to allocate service to all customers between 1 and N'.

It is easy to see the nonmonotonicity of the revenue curve and the impact of Observation 1 by considering the hypothetical bids received for a service as in Table 1.

For example, in Table 1 it is not in the provider's best interest to offer service to all customers between 1 and 8. It is in the provider's interest to provide service to Customers 1, 3, 4, 5, 6, or 8—those that have a nonnegative marginal revenue contribution. The service provider will never choose to provide service to only

Table 1 Example to Demonstrate the Effect of Bundling

Customer	1	2	3	4	5	6	7	8	9	10
Bids ($)	12	10	4.5	4	3.5	3	2.5	2	1.9	1
Revenue	10	9	12	14	15	15	14	15.2	9	
Marginal revenue	10	**−1**	**2**	2	1	**0**	**−1**	**0.2**	−6.2	
Bundle MR	10	**2**		2	1	**0.2**				

the first two or the first seven customers, even if there is available capacity. In these cases the provider's revenue is less than that obtained from providing service to just one or six customers, respectively. An alternative way to characterize this property is that Customer 2 will receive service if and only if Customer 3 receives the service, and Customer 7 will receive the service if and only if Customer 8 receives the service. This motivates us to introduce a simple and intuitive approach that *bundles* the critical customers together when some customers cannot independently receive a service because of a given bid structure. Table 1 shows the two bundles that could be created for this example and the associated marginal revenues after creating these bundles.

Customer bundling implies that either all of the bundled customers are considered for service provision or none. Note that bundling customers does not change the characteristics of the optimal solution. On the other hand, bundling customers provides a way to create a marginal revenue structure such that as soon as the first negative marginal revenue is encountered, a simple scanning algorithm can stop and find the optimal solution.

We define the type of customer bundles created in Table 1 as *the minimal cardinality bundle*. It is a bundle that starts at the first customer index, where the customer's marginal revenue became less than or equal to zero, and ends at the first customer index, where the marginal revenue becomes positive. Note that such bundles can be created as a part of data preprocessing.

Proposition 3 proves that minimum cardinality bundling is an optimal (revenue-maximizing) bundling scheme for a single service class.

Proposition 3. *A bundling strategy that creates a minimal cardinality bundle is optimal for a single service class.*

Proof. Let $R_i = i \cdot B_{i+1}$ be the revenue if customer i is provided a service, and $Rmax_{i-1} = \max\{R_k \mid k < i)$ be the maximum revenue if only the customers having index less than i are considered. Then,

$$mr_i = R_i - Rmax_{i-1}. \tag{19}$$

mr_i is the marginal revenue if customer i is provided the service. Note that as long as $mr_{i-1} > 0$, $Rmax_{i-1} = R_{i-1}$.

Now, consider a stream of customers with marginal revenues mr_{i-1}, mr_i, mr_{i+1}, $\ldots$, mr_{i+k}, $mr_{i+(k+1)}$ such that without loss of generality $mr_i, \ldots, mr_{i+(k-1)} < 0$ and mr_{i-1}, mr_{i+k}, and $mr_{i+(k+1)} > 0$.

In this case, the minimum cardinality bundle will be the one that includes customer $(i+k)$. Let this bundle be called (B1). The marginal revenue for bundle (B1) will be

$$mr_{B1} = mr_{i+k} = R_{i+k} - R_{i-1} > 0. \tag{20}$$

Equation (19) arises from the fact that because $mr_i, \ldots, mr_{i+(k-1)} < 0$, $Rmax_{i+(k-1)} = R_{i-1}$.

Now, consider bundle (B2), which not only includes customer $(i+k)$ but also customer $(i+k+1)$. Because $mr_{i+(k+1)} > 0$ and customer $(i+k)$ is already part of the bundle, the marginal revenue for bundle B2 will be

$$mr_{B2} = (R_{i+(k+1)} - R_{i-1}) > mr_{B1}. \tag{21}$$

Even though by including the customer $(i+k+1)$ we may increase the marginal revenue, it could only hurt the maximum revenue by excluding the whole bundle due to the higher capacity required. To see this we need consider only two cases:

(i) *When there is enough capacity to include customer* $(i+k+1)$. In this case, after including bundle B1, the customer $(i+k+1)$ will automatically be included since $mr_{i+(k+1)} > 0$. Therefore bundles (B1) and (B2) both will provide identical revenue R_{i+k+1}.

(ii) *When there is not enough capacity to include customer* $(i+k+1)$ *but enough to include* $(i+k)$. In this case if we create bundle (B2), then due to capacity restrictions, the whole bundle could not be considered and the revenue would be R_{i-1}. However, bundle (B1) can be included and the revenue would be $R_{i+k} > R_{i-1}$.

Therefore B2 is a suboptimal bundle and the optimal bundle size is the minimal cardinality bundle. Q.E.D.

Thus, under the MVA, a seller has to consider the fact that every additional bidder's inclusion can have a potentially negative impact on the price of all prior accepted bidders.

2.3.5. First-Price "Pay-Your-Bid" Auction. For the sake of completeness, we present the properties of a first-price pay-your-bid scheme in the context of our problem. It is well known that such a scheme induces bid-shading (Krishna 2002) and bidders will not have the incentive to truth-tell. Interestingly, it also removes the structural nicety of the MVA where a seller was forced to accept all higher bids in a class if she was making a move to accept a given bid. Thus, we don't have the pseudo-polynomial-complete enumeration computational nicety that we had in the MVA and the mMVA. One could make a case for the cognitive simplicity of such a mechanism, but the billions of second-price eBay auctions that have been conducted challenge the viability of such an argument.

Table 2 Comparison of Auction Schemes—MVA Balances Computational Efficiency and Posterior Regret

Auction scheme	Problem complexity	Computational complexity	Economic properties
GVA	NP-hard Not suitable for primal-dual algorithms Formulation requires the enumeration of all feasible solutions	Significant, as it requires solution of $N+1$ NP-hard problems to determine transfers	Allocatively efficient Incentive compatible No commodity prices.
MVA	NP-hard	Low, special structure lends to pseudo-polynomial enumeration. Alternatively greedy heuristics can be used to get ε-optimal results	Revenue maximizing. Not incentive compatible due to variable supply in revenue maximization context. However, posterior regret is bounded. Imputed commodity prices
Max MVA	NP-hard	Same as MVA	Efficient Incentive compatible Nonrevenue maximizing Imputed commodity prices Perceived fair.
First price pay your bid	NP-hard knapsack with no special structure	Only rely on greedy algorithms. Worst case performance known to be bad.	Not incentive compatible Not necessarily efficient due to bid-shading No imputed prices Easy to understand

3. Summary of Findings and Future Research

This paper contributes by coupling auction-based preference-elicitation schemes for information goods with knapsack-based resource allocation and pricing. We do this in the context of optimizing Internet content providers' profits in the presence of unknown demand and multiple service classes. We formulate and contrast a portfolio of preference revelation and pricing and allocation schemes with respect to their theoretical and computational complexity, as well as their economic properties. Table 2 summarizes the comparative features of the four auction mechanisms we covered.

We recommend use of the GVA for service providers who either have in-house combinatorial optimization expertise or can acquire such expertise for a net gain and who do not necessarily need real-time solutions. For the vast majority of commercial real-world sellers who prefer to have commodity prices, we expect that the MVA offers a nice compromise between computational and economic properties. In addition, for social welfare–maximizing settings, we would recommend that service providers use the mMVA. We would not recommend the first-price scheme to any seller. In this research we have not considered the consumer bidding strategies explicitly. In future research we will address such issues by studying the effect of various design modifications on consumer bidding strategies. Several more interesting questions arise in such markets that need to be explored. Should the provider adopt a clearing-house approach and make accept/reject decisions at regular intervals? Should the reservation price be altered each day in such scenarios? Alternatively, would traditional yield management strategies, such as those used by airlines and hotels, enable a provider to make a probabilistic assessment of future bid patterns and consequently make instant accept/reject decisions?

Acknowledgments

Author names are listed in alphabetical order. Alok Gupta's research is supported by NSF CAREER Grant IIS-0301239 and IIS-0219825, but does not necessarily reflect the views of the NSF. Partial support for this research was also provided by CIDRIS and TECI—the Treibick Electronic Commerce Initiative, OPIM/SBA, University of Connecticut.

References

Ahuja, R., J. B. Orlin, A. Tiwari. 2000. A greedy genetic algorithm for the quadratic assignment problem. *Comput. Oper. Res.* **27** 917–934.

Ausubel, L. M. 2004. An efficient ascending bid auction for multiple objects. *Amer. Econom. Rev.* **94**(5) 1452–1475.

Bapna, R., P. B. Goes, A. Gupta. 2000. A theoretical and empirical investigation of multi-item online auctions. *Inform. Tech. Management* **1**(1) 1–23.

Bapna, R., P. B. Goes, A. Gupta. 2003. Analysis and design of business-to-consumer online auctions. *Management Sci.* **49**(1) 85–101.

Bikhchandani, S., J. W. Mamer. 1997. Competitive equilibrium in an exchange economy with indivisibilities. *J. Econom. Theory* **74** 385–413.

Bikhchandani, S., J. Ostroy. 2002. The package assignment model. *J. Econom. Theory* **107** 377–406.

Clarke, E. H. 1971. Multipart pricing of public goods. *Public Choice* **11** 17–33.

de Vries, S., R. Vohra. 2003. Combinatorial auctions: A survey. *INFORMS J. Comput.* **15**(3) 284–309.

Foley, D. 1967. Resource allocation and the public sector. *Yale Econom. Essays* **7** 45–98.

Greenwood, J., R. P. McAfee. 1991. Externalities and asymmetric information. *Quart. J. Econom.* **106**(1) 103–121.

Groves, T. 1973. Incentives in teams. *Econometrica* **41** 617–631.

Gupta, A., D. O. Stahl, A. B. Whinston. 1997. A stochastic equilibrium model of Internet pricing. *J. Econom. Dynam. Control* **21** 697–722.

Krishna, V. 2002. *Auction Theory*. Academic Press, San Diego, CA.

Lazar, A. A., N. Semret. 2000. Design and analysis of the progressive second price auction for network bandwidth sharing. *Telecomm. Systems* http://comet.columbia.edu/~nemo/pub.html.

List, John A., David Lucking-Reiley. 2000. Demand reduction in multi-unit autions: Evidence from a sportscard field experiment. *Amer. Econom. Rev.* **90**(4) 961–972.

Lucking-Reiley, D. 2000. Auctions on the Internet: What's being auctioned, and how? *J. Indust. Econom.* **48**(3) 227–252.

MacKie-Mason, J., H. Varian. 1995. Pricing the Internet. B. Kahin, J. Keller, eds. *Public Access to the Internet.* Prentice-Hall, Englewood Cliffs, NJ.

Martello, S., P. Toth. 1989. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley and Sons, Chichester, England, 27–29.

O'Hara, M. 1995. *Market Microstructure Theory*. Blackwell Publishers, Cambridge, MA.

RealSystem Server Best Practices. 2001. RealSystem, http://docs.real.com/docs/dg.pdf

Vickrey, W. 1961. Counter-speculation, auctions, and competitive sealed tenders. *J. Finance* **41** 8–37.

Wang, R. 1993. Auctions versus posted-price selling. *Amer. Econom. Rev.* **84**(4) 838–851.