Strategy-Proof Data Auctions with Negative Externalities

Xiang Wang†, Zhenzhe Zheng‡, Fan Wu†, Xiaojie Tang†, and Guihai Chen†
†Shanghai Key Laboratory of Scalable Computing and Systems
Department of Computer Science and Engineering
Shanghai Jiao Tong University, China
‡Department of Information Systems, University of Texas at Dallas, USA

ABSTRACT
Data has appeared to be a new kind of commodity with distinctive characteristics, which make it fundamentally different from physical goods as well as traditional digital goods. Therefore, new trading mechanisms for data need to be designed. In this paper, we model the data market as an auction with negative externalities, and design practical mechanisms for data trading. Specifically, we propose a family of Data auctions in Competitive markets, namely DICTA. DICTA contains two mechanisms, including DICTA-FUL and DICTA-PAR. DICTA-FUL is a direct revelation auction mechanism in full competition markets, achieving strategy-proofness and optimal social welfare. In the partial competition markets, we show that the allocation problem is NP-hard. Therefore, we present approximation algorithm and a charging scheme, DICTA-PAR achieves both strategy-proofness and d-approximation, where d is the maximum degree of the underlying undirected graph of the competition graph. Finally, we use both artificial and real (crawled from Google Finance) competition relations to evaluate our auction mechanisms. Evaluation results show that our auction mechanisms achieve good performance in terms of social welfare, satisfaction ratio, and revenue.

Categories and Subject Descriptors
J.4 [Computer Applications]: Social and Behavioral Science—Economics

General Terms
Algorithms, Theory, Economics

Keywords
Data Market, Auction, Externality

1. INTRODUCTION
In recent years, data has become a new kind of commodity that can be traded on the Internet. For example, Xignite sells financial data, Gnip sells data from social networks, and Sabre trades consumers’ booking and searching data on travel. To facilitate online data marketing, several platforms have emerged, e.g., Azure Data Marketplace, Infochimps, and Dataexchange. These centralized platforms let data owners upload and sell their data, and let data consumers discover and purchase the data needed.

However, data as a kind of commodity is fundamentally different from physical goods, since data exhibits a distinctive characteristic, i.e., once produced, the data can be duplicated for any number of copies with low or no cost [47]. A related but different kind of commodities in online markets is digital goods, such as electronic books, audio files, and pay-per-view movies. Although extensive studies have been carried out on pricing of digital goods, they cannot be directly applied to trading of data. In digital good auctions, the negative externalities do not exist. However, data buyers may want to possess the data exclusively or to limit the distribution of data copies to their competitors. According to a recent survey on data market [46], out of all vendors in the research, 87% of offered data is in business contexts. Buyers, who are mostly companies, purchase their interest-ed data in order to gain advantages in their business against their competitors. Such advantages can be undermined, if their competitors also get the same data, called negative externalities. Therefore, a buyer’s valuation on the data not only depends on whether she can get the data set, but also on the data allocation to her competitors.

Since strategy-proof auctions can effectively elicit buyer-s’ true valuation and distribute goods efficiently, we seek to design effective data auctions in this preliminary work. However, it is not a trivial job to achieve this goal. The first challenge is how to prevent rational and selfish buyers from manipulating the auction. Although the design of strategy-proof (Please refer to Section 2 for the definition) single-parameter auctions has been well investigated, existing approaches cannot be applied to data auctions, because there are three private parameters for each of the buyers, including valuation, set of competitors, and the maximum number of her competitors that can obtain the data at the same time without diminishing her valuation to zero (namely tolerance bound). Multi-parameter mechanism design is still an open question.

Yet, another challenge is on the problem of social welfare maximization. In general data markets, the buyers may have different sets of competitors, making the searching for the allocation achieving the optimal social welfare computationally intractable.

In this paper, we conduct an in-depth study on designing strategy-proof data auctions. We propose a family of Data auction in Competitive markets (DICTA). DICTA contains two mechanisms, namely DICTA-FUL and DICTA-PAR. Specifically, DICTA-FUL is for an ideal but meaningless setting, i.e., full competition markets, where any pair of buyers compete with each other. In this scenario, we propose
a computationally tractable algorithm to calculate an optimal allocation, so that Vickrey-Clarke-Groves (VCG) mechanism [49, 17, 25] can be applied to achieve both efficiency and strategy-proofness. We further consider a more practical scenario, i.e., partial competition markets, in which each buyer just competes with a subset of the buyers instead of all. In this setting, the VCG mechanism is no longer appropriate due to computational intractability of searching for the optimal allocation. Therefore, we turn to design an approximation algorithm that finds a sub-optimal allocation, and incorporate it with a simple but effective pricing scheme to achieve strategy-proofness.

The major contributions of this paper are summarized as follows.

- To the best of our knowledge, we are the first to model the data trading market as an auction considering the distinctive characteristics of data as a kind of commodity and mutually exclusive competitions among the buyers.

- For full competition data markets, we present DICTA-FUL to achieve both optimal social welfare and strategy-proofness.

- For partial competition data markets, considering the computational intractability of locating the optimal allocation, we propose an approximately optimal data auction mechanism DICTA-PAR achieving strategy-proofness and an approximation ratio of \(d\), where \(d\) is the maximal degree of the underlying undirected and simple graph of competition graph (Please refer to Section 4.2 for the definition).

- We have extensively evaluated DICTA-FUL and DICTA-PAR using both artificial and real competition relations, respectively. Especially for DICTA-PAR, the competition relations are crawled from different market sectors on Google Finance. Evaluation results show that our auction mechanisms achieve good performance, in terms of social welfare, satisfaction ratio, and revenue.

The rest of the paper is organized as follows. In Section 2, we present the model of data auction considering competition between buyers. In Section 3, we propose DICTA-FUL for the full competition markets. We further consider partial competition markets, and propose DICTA-PAR in Section 4. In Section 5, we show the evaluation results. In Section 6, we concisely review related works. Finally, we conclude the paper in Section 7.

## 2. MODEL OF DATA AUCTION

In this section, we present the auction model of a data market, and then review the most closely related solution concepts from algorithmic mechanism design.

We consider a one shot sealed bid data auction with a trusted auctioneer and a set of \(n\) buyers \(N = \{1, 2, 3, \ldots, n\}\). There is a single set of data that can be duplicated to any number of copies, and then be sold to different interested buyers. Each buyer \(i \in N\) is interested in a single copy of the data set.

The existence of externalities in auctions is due to the auctioned good might play a role in future interactions among the auction’s participants. So the outcome of the auction affects the future interaction, and thus a buyer is no longer indifferent about the identity of the winner of the auction. Specifically, in data auctions, the winner can exploit the information resource in the set of data and gain advantages against their competitors. However, such advantages can be undermined, if their competitors also get the same data. So a buyer is highly motivated to limit the distribution of data copies to their competitors.

In order to generally express the externalities, specifying a valuation in a single-item data auction of \(n\) buyers requires providing a value for each of the \(2^n\) allocation results, which is highly impractical. So in this preliminary work, we adopt a simple but effective binary valuation. Each buyer has a set of business competitors \(S_i \subseteq N \setminus \{i\}\), and can only tolerate up to \(t_i\) competitors sharing the same set of data, i.e., winning the set of data in the data auction. Otherwise, the buyer may lose the advantage by having the data set in her business. If the buyer \(i\) wins the data set and there are no more than \(t_i\) competitors winning at the same time, then she has a valuation \(v_i\) on the data set. Otherwise, i.e., the buyer \(i\) loses the auction or more than \(t_i\) competitors win the auction, the valuation of the data set to the buyer \(i\) becomes 0. Suppose the winner set is \(W\), we define the valuation function of each buyer \(i\) as follows:

\[
v_i(W) = \begin{cases} 
  v_i & i \in W : |W \cap S_i| \leq t_i, \\
  0 & \text{otherwise.}
\end{cases}
\]  

The triple \(\theta_i = (S_i, t_i, v_i)\) is the private information of buyer \(i\), and is widely known as type in the literature. We consider that the buyers are rational and selfish, and thus may misreport their sets of competitors, tolerance bounds, and valuations, if doing so can increase their payoffs/utilities. In the data auction, each buyer \(i\) proposes a bid \(b_i = (\hat{S}_i, \hat{t}_i, \hat{v}_i)\), which can differ from her type. We denote the bidding profile of all the buyers as \(\hat{b} = (b_1, b_2, \cdots, b_n)\). The primary design objective of this work is to prevent misbehaviors of the buyers, such that bidding one’s type is the best strategy to each buyer.

After collecting the bids from the buyers, the auctioneer construct a directed competition graph \(G\), in which each vertex represents a buyer, and each edge \((i, j), i, j \in N\) indicates that buyer \(j\) is in buyer \(i\)’s competitor set. Then, the auctioneer determines a set of winners \(W \subseteq N\), to each of whom a copy of the data set is distributed. The auctioneer also calculates a payment \(p_i\) for each buyer \(i \in W\). The buyers are free of charge, if they lose the auction.

We now can define the utility of each buyer \(i\), which is the difference between her valuation and payment as follows.

\[
u_i = \begin{cases} 
  v_i - p_i, & i \in W : |W \cap S_i| \leq t_i, \\
  -p_i, & i \in W : |W \cap S_i| > t_i, \\
  0, & \text{otherwise.}
\end{cases}
\]  

From the perspective of auctioneer, the design objective of the auction mechanism is to maximize social welfare, which is the sum of the winners’ valuations on the allocated data set.
\[ SW = \sum_{i \in \mathcal{W}} v_i(\mathcal{W}) = \sum_{i \in \mathcal{W}, |\mathcal{W} \cap \mathcal{S}_i| \leq t_i} v_i. \]  

(3)

The perfectly matching solution concept with our design objectives is strategy-proofness from algorithmic mechanism design.

**Definition 1.** A direct revelation mechanism is strategy-proof, if it satisfies both dominant strategy incentive-compatibility and individual-rationality.

Here, dominant strategy incentive-compatibility means that there is no buyer having incentive to cheat about her private information (type), giving that revealing true information is the best strategy to herself, which maximizes her utility no matter what the others do. The complementary concept, individual-rationality, means that every buyer truthfully revealing her type in the game never gets a negative utility.

Therefore, the objective of this work can be stated as to design strategy-proof and computationally efficient data auction mechanisms to maximize social welfare in the data auction.

In this paper, we consider two progressive scenarios, including full competition markets (Section 3) and partial competition markets (Section 4). In the full competition markets, each pair of buyers are competitive to each other, and thus forming a complete competition graph. Partial competition market is a more general case, in which each buyer just competes with a subset of the buyers instead of all.

### 3. DATA AUCTION IN FULL COMPETITION MARKETS

In this section, we present data auction mechanism DICTA-FUL for the full competition markets, in which any pair of buyers compete with each other, i.e., the competition graph is a complete graph.

In this scenario, we can design a polynomial time algorithm to compute the optimal allocation, and thus can apply the celebrated Vickrey-Clarke-Groves (VCG) mechanism [40, 17, 25] to achieve strategy-proofness. Therefore, we focus on algorithm design of allocation rule in this section.

DICTA-FUL consists of a data set allocation rule and a payment determination rule.

#### 3.1 Data Set Allocation

We present our computationally efficient algorithm for data set allocation, and prove its optimality in terms of social welfare.

In the data set allocation algorithm, we first sort all the buyers in a non-increasing order of their declared valuations, and denote the sorted list by \( \Gamma \):

\[ \Gamma : \hat{v}_1 \geq \hat{v}_2 \geq \cdots \geq \hat{v}_n. \]

If there exists a tie, we break it arbitrarily. We note that \( \hat{v}_i \) may not be equal to \( v_i \) after sorting, and we will apply the allocation algorithm to the buyers according to the order in the sorted list.

Since any pair of buyers compete with each other in the full competition markets, for each winner, the number of her winning competitors is equal to the number of all the winners minus one. Thus, we can traverse every possible number of winners from 1 to \( n \). For each number \( m \in \{1, 2, \ldots, n\} \), we pick top \( m \) buyers whose tolerance bounds are no less than \( m - 1 \) from the sorted list \( \Gamma \). Therefore, we focus on algorithm design of allocation rule in this section.

**Algorithm 1** Data set allocation algorithm of DICTA-FUL: Alloc-FUL\((\mathbb{N}, \hat{b})\)

**Require:** A set of buyers \( \mathbb{N} \) and their bidding profile \( \hat{b} \);

**Ensure:** A set of winners \( \mathcal{W} \);

1. Sort the buyers in non-increasing order of their bidding values: \( \Gamma : \hat{v}_1 \geq \hat{v}_2 \geq \cdots \geq \hat{v}_n \);
2. \( \mathcal{W} \leftarrow \emptyset \);
3. for \( m = 1 \) to \( n \) do
5. for \( i = 1 \) to \( n \) do
6. if \( |\mathcal{W}| < m \) then
7. if \( \hat{v}_i \geq m - 1 \) then
8. \( \mathcal{W}' \leftarrow \mathcal{W} \cup \{i\} \);
9. end if
10. else
11. break;
12. end if;
13. end for
14. if \( \sum_{i \in \mathcal{W}'} \hat{v}_i > \sum_{i \in \mathcal{W}} \hat{v}_i \) then
16. end if
17. end for
18. return \( \mathcal{W} \).

**Theorem 1.** DICTA-FUL achieves optimal social welfare.

**Proof.** We prove the theorem by contradiction. Suppose that there exist a set \( \mathcal{W}' \subset \mathbb{N} \) of winners, which achieves a higher social welfare than \( \mathcal{W} \) calculated by DICTA-FUL. We denote the social welfare of \( \mathcal{W} \) and \( \mathcal{W}' \) by \( SW \) and \( SW' \), respectively.

We first prune the set \( \mathcal{W}' \) to eliminate possibly invalid winners. If there exists any buyer \( i \in \mathcal{W}' \), such that \( t_i < |\mathcal{W}'| - 1 \), we can remove it from the set \( \mathcal{W}' \) without decreasing the social welfare, because her valuation is 0 in this case. For convenience of notation, we still use \( \mathcal{W}' \) after pruning.

Let \( k = |\mathcal{W}'| \). We denote the winner set and the corresponding social welfare in Algorithm 1 in the \( k \)th iteration by \( \mathcal{W}_k \) and \( SW_k \), respectively. Given \( \mathcal{W}' \), there must be at least \( k \) buyers with tolerance bounds no less than \( k - 1 \). Consequently, \( |\mathcal{W}_k| = |\mathcal{W}'| \). Since Algorithm 1 always selects top \( k \) buyers by valuations without violating their tolerance bounds, we have \( SW_k \geq SW' \).

Furthermore, the final winner set and its corresponding social welfare is picked among all the \( n \) possible numbers of winners, i.e.,

\[ SW = \max\{SW_m|m \in \{1, 2, \ldots, n\}\} \geq SW_k \geq SW'. \]
Here comes the contradiction. Therefore, there exist no other allocations, whose social welfare exceed that achieved by DICTA-FUL. □

3.2 Payment Determination

Considering that the data set allocation rule selects the winner set with optimal social welfare, we can apply the celebrated VCG mechanism to determine the payment for each of the winners. In particular, the payment $p_i$ of each buyer $i \in \mathbb{N}$ is calculated as follows.

$$p_i = \begin{cases} \sum_{j \in \text{Alloc-FUL}(i \cup \{i\}, \vec{b})} \hat{v}_j - \sum_{j \in \text{W} \setminus \{i\}} \hat{v}_j, & i \in \text{W}, \\ 0, & \text{otherwise.} \end{cases}$$

(4)

Here, $\text{Alloc-FUL}(\mathbb{N} \setminus \{i\}, \vec{b})$ is the winner set for buyers without $i$, and $\hat{v}_j$ is in the sorted list $\Gamma'$ for the buyers without $i$. According to the design principles of the VCG mechanism, we can claim the following theorem.

THEOREM 2. DICTA-FUL is a strategy-proof auction mechanism.

4. DATA AUCTION IN PARTIAL COMPETITION MARKETS

In this section, we further consider the design of data auction in partial competition markets. In partial competition markets, each buyer competes with a subset of the buyers. The previously studied full competition markets are special cases of the partial competition markets. The model of partial competition markets is more general and practical, because it captures the essence of the data set valuable to companies, which belong to different market sectors, but only need to compete with the ones in the same sector. In this section, we will first show the hardness of allocation in this general scenario, and then present a novel approximation allocation algorithm combined with a payment scheme based on the concept of critical bid. The auction mechanism presented in this section is called DICTA-PAR.

4.1 Hardness of the Allocation Problem

Unfortunately, having less competitions among the buyers makes the problem of winner selection achieving optimal social welfare computationally intractable. We show that the allocation problem in partial competition markets is NP-hard, and also hard to approximate. We first prove this problem is NP-hard by reduction from the independent set problem as follows:

THEOREM 3. The allocation problem in partial competition markets is NP-hard.

PROOF. We will make a reduction from the NP-complete independent set problem: given an undirected graph $G = (V, E)$ and a number $k$, does $G$ have an independent set of size $k$? Given such an independent set instance, we will build an allocation problem from it as follows: Each vertex in the graph represents a buyer. For vertex $i \in V$, we will have the competitor set of $i$ be the set of adjacent vertices $S_i = \{v \in V \mid (v, i) \in E\}$. For every buyer $i$, its tolerance bound $t_i$ is 0 and value $v_i$ is 1.

We note that a set $\text{W}$ of winners in the allocation problem satisfies $i \in \text{W} : |\text{W} \cap S_i| \leq t_i$ if and only if the set of vertices corresponding to $W$ is an independent set in the graph $G$. The social welfare obtained by $W$ is exactly the size of the independent set. So an independent set of size at least $k$ exists if and only if the social welfare of the optimal allocation is at least $k$, which concludes the NP-hardness proof. □

Also, it is known that approximating independent set problem within a factor of $n^{1-\epsilon}$ (for any fixed $\epsilon > 0$) is NP-complete. Since in our reduction, the social welfare was exactly equal to the size of the independent set, we conclude as follows:

THEOREM 4. Approximating the optimal allocation in partial competition markets within a factor better than $n^{1-\epsilon}$ (for any fixed $\epsilon > 0$) is NP-hard.

Since the computational intractability of computing optimal allocation, VCG mechanism cannot be applied here. Furthermore, the above theorem rules out the possibility of finding an allocation algorithm with any constant approximation factor. In the next subsection, we will show an allocation algorithm with approximation $d$, which is the maximum degree of the underlying undirected and simple graph of the competition graph.

4.2 Data Set Allocation

Algorithm 2 Data set allocation algorithm in DICTA-PAR: Alloc-PAR($\mathbb{N}, \vec{b}$)

Require: A set of buyers $\mathbb{N}$ and their bidding profile $\vec{b}$;
Ensure: A set of winners $\text{W}$;

1: Sort the buyers in non-increasing order of their bidding values: $\Gamma : \hat{v}_1 \geq \hat{v}_2 \geq \cdots \geq \hat{v}_n$;
2: $\text{W} \leftarrow \emptyset$;
3: for $i = 1$ to $n$ do
4: \hspace{1em} if $\left(\forall j \in \{k | k \in \text{W} \wedge i \in S'_k\}, \hat{v}_j > |S'_i \cap \text{W}|\right)$ \& $(\hat{v}_i \geq |S'_i \cap \text{W}|)$ then
5: \hspace{2em} $\text{W} \leftarrow \text{W} \cup \{i\}$;
6: \hspace{1em} end if
7: \hspace{1em} end for
8: return $\text{W}$.

We here present our computationally efficient algorithm for data set allocation achieving $d$-approximation ratio. Same as before, we first sort all the buyers in a non-increasing order of their declared valuations, and denote the sorted list by $\Gamma$.

$$\Gamma : \hat{v}_1 \geq \hat{v}_2 \geq \cdots \geq \hat{v}_n.$$ If there exists a tie, we break it arbitrarily.

Following the sequence specified in $\Gamma$, we visit each buyer one by one, and check whether she can be allocated the data set without violating the following two constraints. ▶ The first constraint is that allocating the data set to buyer $i$ should not breach any of the previously selected winners’ tolerance bounds, i.e.,

$$\forall j \in \{k | k \in \text{W} \wedge i \in S'_k\}, \hat{v}_j > |S'_i \cap \text{W}|.$$ ▶ The second constraint is that the number of previously selected winning competitors of buyer $i$ should not exceed
her tolerance bound, i.e.,
\[ i' \geq \left| S'_i \cap \mathcal{W} \right| . \]

If both the above constraints are satisfied, we allocate the data set to buyer \( i \); Otherwise, we deny buyer \( i \)'s bid.

Algorithm 2 shows the pseudo-codes of DICTA-PAR’s allocation rule. The complexity of Algorithm 2 is \( O(n^2) \).

\[ \begin{aligned}
\text{Algorithm 2: DICTA-PAR's allocation rule.}
\end{aligned} \]

**Figure 1:** Transform the competition graph to its underlying undirected graph.

Next, we show the definition of \( d \), and then prove the approximation ratio of DICTA-PAR is \( d \).

As Figure 1 illustrates, we construct the underlying undirected and simple graph \( G' \) of competition graph \( G \) by replacing all directed edges of \( G \) with undirected edges and remove the redundant edge (if exists) between any two vertices. We denote the maximum degree of \( G' \) by \( d \).

**Theorem 5.** DICTA-PAR’s allocation rule achieves an approximation ratio of \( d \).

**Proof.** Let \( \mathcal{O} \subseteq \mathbb{N} \) be the optimal data allocation, and \( \mathcal{W} \subseteq \mathbb{N} \) be the allocation computed by DICTA-PAR. Since DICTA-PAR achieves strategy-proofness, we can use the types of the buyers as bids here.

If buyer \( j \in \mathbb{N} \setminus \mathcal{W} \) can win the auction when buyer \( i \in \mathcal{W} \) is removed from \( \mathbb{N} \) before auction, we say that buyer \( i \) prevents buyer \( j \) from being allocated. For each buyer \( j \in \mathcal{O} \setminus \mathcal{W} \), let \( i \in \mathcal{W} \setminus \mathcal{O} \) be one of the buyers selected by DICTA-PAR preventing buyer \( j \) from being allocated. Then, we distinguish three cases in the following analysis.

- **Case 1:** Buyer \( j \) is one of buyer \( i \)'s competitors, i.e., \( j \in S_i \).
- **Case 2:** Buyer \( i \) is one of buyer \( j \)'s competitors, i.e., \( i \in S_j \).
- **Case 3:** Buyers \( j \) and \( i \) both are another buyer \( k \)'s competitors. We limit \( k \in \mathcal{W} \cap \mathcal{O} \) here, since the case, where \( k \in \mathcal{W} \setminus \mathcal{O} \), coincides with Case 1.

For each buyer \( i \in \mathcal{W} \setminus \mathcal{O} \), we define
\[ \mathcal{O}_i \triangleq \left\{ j \mid j \in \mathcal{O} \setminus \mathcal{W} \land v_j \leq v_i \land (j \in S_i \lor i \in S_j) \right\} \] (5)
to represent the buyers in \( \mathcal{O} \) that cannot be allocated data by DICTA-PAR, possibly due to conflict with winning buyer \( i \) in Case 1 or Case 2.

In the Case 3, we note that for each buyer \( i \in \mathcal{W} \setminus \mathcal{O} \) and each buyer \( k \in \mathcal{W} \cap \mathcal{O} \), buyer \( i \) can prevent at most one buyer \( j \in \mathcal{O} \setminus \mathcal{W} \), \( i, j \in S_k \). Because the removal of buyer \( i \) can only vacate one room for \( k \)'s competitors, then permits at most one buyer in \( \mathcal{O} \setminus \mathcal{W} \) to win. So the number of buyers in \( \mathcal{O} \setminus \mathcal{W} \) prevented by buyer \( i \) is no more than \( |\{ k \mid k \in \mathcal{O} \cap \mathcal{W} \land i \in S_k \}| \), denoted by \( m_i \).

For each buyer \( i \in \mathcal{W} \setminus \mathcal{O} \), we define \( \mathcal{M}_i \) to represent the buyers in \( \mathcal{O} \) that cannot be allocated data by DICTA-PAR, possibly due to conflict with winning buyer \( i \) in Case 3. It is evident that \( v_j \leq v_i, \forall j \in \mathcal{M}_i \).

Given the maximum degree \( d \) of the underlying undirected and simple graph of competition graph, there are at most \( d - m_i \) buyers in \( \mathcal{O}_i \). Then, we have
\[ \sum_{j \in \mathcal{O}_i} v_j \leq (d - m_i) \times v_i. \] (6)

Similarly, we have
\[ \sum_{j \in \mathcal{M}_i} v_j \leq m_i \times v_i. \] (7)

Since \( \mathcal{O} \setminus \mathcal{W} \subseteq \bigcup_{i \in \mathcal{W} \setminus \mathcal{O}} (\mathcal{O}_i \cup \mathcal{M}_i) \), we get
\[ \sum_{i \in \mathcal{O} \setminus \mathcal{W}} v_i \leq \sum_{i \in \mathcal{W} \setminus \mathcal{O}} \left( \sum_{j \in \mathcal{O}_i} v_j + \sum_{j \in \mathcal{M}_i} v_j \right) \]
\[ \leq (d - m_i + m_i) \times \sum_{i \in \mathcal{W} \setminus \mathcal{O}} v_i \]
\[ = d \times \sum_{i \in \mathcal{W} \setminus \mathcal{O}} v_i \] (8)

Since \( \mathcal{O} = (\mathcal{O} \setminus \mathcal{W}) \cup (\mathcal{W} \cap \mathcal{O}) \) and \( \mathcal{W} = (\mathcal{W} \setminus \mathcal{O}) \cup (\mathcal{W} \cap \mathcal{O}) \), we have
\[ \sum_{i \in \mathcal{O}} v_i = \sum_{i \in \mathcal{W} \setminus \mathcal{O}} v_i + \sum_{i \in \mathcal{W} \cap \mathcal{O}} v_i \]
\[ \leq \sum_{i \in \mathcal{W} \setminus \mathcal{O}} v_i + d \times \sum_{i \in \mathcal{W} \setminus \mathcal{O}} v_i \]
\[ \leq d \times \left( \sum_{i \in \mathcal{W} \setminus \mathcal{O}} v_i + \sum_{i \in \mathcal{W} \cap \mathcal{O}} v_i \right) \]
\[ = d \times \sum_{i \in \mathcal{W}} v_i. \] (9)

Therefore, the approximation ratio of DICTA-PAR is \( d \). \( \square \)

We note that the approximation ratio of DICTA-PAR’s allocation rule degrades to \( n - 1 \) in the worst case. Unfortunately, considering the difficulties in designing strategy-proof auctions with multi-parameter bids, this is the best algorithm we can come up with in this preliminary work. However, we will keep on designing strategy-proof auctions with lower approximation ratios in our future work.

### 4.3 Payment Determination

Since DICTA-PAR follows a greedy allocation rule, we may adopt the concept of critical bid to determine the payment for each of the buyers. We have to note that normally it is not appropriate to apply critical bid to design strategy-proof mechanisms for triple-parameter auctions. Given the distinctive characteristics of the binary valuation function in our data auction, we can carefully eliminate the possibility of cheating one’s tolerance bound and set of competitors, and thus achieve strategy-proofness by adopting critical bid.

**Definition 2 (Critical Bid).** The critical bid of a winner is the minimum valuation she should declare to win the auction.

The critical bid of each winner \( i \in \mathcal{W} \) can be calculated by re-executing Algorithm 2 to find the first buyer that has
been denied, but would have been allocated the data set if we do not allocate the data set to buyer $i$. We denote the critical buyer of $i$ by $\pi(i)$.

We can now define the payment $p_i$ of each buyer $i$:

$$p_i = \begin{cases} 
\hat{v}_{\pi(i)}', & i \in W \land \pi(i) \text{ exists}, \\
0, & \text{otherwise}.
\end{cases}$$  \hfill (10)

Before proving DICTA-PAR’s strategy-proofness, we first show an important property of DICTA-PAR, which is also the essential for the later proof of strategy-proofness.

**Lemma 1.** In DICTA-PAR, no buyer can decrease her critical bid without diminishing her valuation to zero.

**Proof.** According to the definition of critical bid, we know that declaring a different valuation cannot manipulate the critical bid. The only way to change critical bid is to misreport one’s tolerance bound or set of competitors.

Considering DICTA-PAR’s allocation rule, a buyer loses in the auction must due to one of the following two reasons. The first reason is that allocating the data set to the buyer would breach one of the previously selected winner’s tolerance bound. The second one is that the number of selected winning competitors of the buyer already exceeds her tolerance bound.

If there is a buyer $i$ who decreases her critical bid, then her true critical bidder $\pi(i)$ must be able to win in the auction. To make this happen, buyer $i$ must lose her tolerance bound or shrink her set of competitors. However, either of the manipulations would lead to the breach of buyer $i$’s tolerance bound, and thus diminishes her valuation to zero. \[\square\]

Then, we can show the strategy-proofness of DICTA-PAR.

**Theorem 6.** DICTA-PAR is a strategy-proof data auction mechanism in partial competition markets.

**Proof.** First, we prove that DICTA-PAR satisfies individual rationality. For each buyer $i$, if she wins the auction and her tolerance bound is not breached, she gets a utility $u_i = v_i - p_i$. Otherwise, her utility is zero. Since truthful bidding is any buyer’s dominate strategy as we shown above, bidding one’s type can lead to a non-negative utility, if she wins the auction.

$$u_i = v_i - p_i = v_i - \hat{v}_{\pi(i)}' \geq 0.$$  \hfill (11)

Next, we prove that DICTA-PAR satisfies incentive compatibility. Suppose that when buyer $i \in N$ bids truthfully, i.e., $b_i = \theta_i = (S_i, t_i, v_i)$, she needs to pay $p_i$ and receives a utility $u_i = v_i - p_i$. Let $p_i'$ and $u_i'$ be buyer $i$’s payment and utility, respectively, when she manipulates her bid, i.e., $b_i' = (S_i', t_i', v_i') \neq \theta_i$. We denote the critical bidder of buyer $i$, when she bids truthfully or not, as $\pi(i)$ and $\pi'(i)$, respectively.

Since buyer $i$ cannot increase her utility if she loses the auction after cheating, we only need to consider the cases, in which buyer $i$ wins the auction after cheating. We distinguish two cases:

- **Buyer $i$ wins the auction when bidding truthfully.** We further distinguish two cases here.
  - Her critical bid decreases after cheating, i.e., $\hat{v}_{\pi'(i)}' < \hat{v}_{\pi(i)}'$. According to Lemma 1, buyer $i$ decreases her critical bid must by breaching her tolerance bound, and thus diminishes her valuation to zero. Then,
    $$u_i' = 0 - p_i' \leq 0 \leq u_i.$$

- **Her critical bid increases after cheating, i.e., $\hat{v}_{\pi'(i)}' \geq \hat{v}_{\pi(i)}'$.** We have
    $$u_i' = v_i - p_i' \leq v_i - \hat{v}_{\pi'(i)}' \leq v_i - \hat{v}_{\pi(i)}' = u_i.$$

So DICTA-PAR achieves strategy-proofness. \[\square\]

5. EVALUATION RESULTS

We have implemented our design of data auctions, including DICTA-FUL and DICTA-PAR. Since we have proven that DICTA-FUL achieves optimal social welfare, we just present DICTA-FUL’s single results in this section. In contrast, we compare DICTA-PAR with a not strategy-proof but approximately optimal solution (denoted by APX-OPT) on social welfare and satisfaction ratio. APX-OPT is computed by solving the integer optimization program using Gurobi 6.0.5 with tolerance $10^{-4}$. Since calculating an APX-OPT allocation is extremely time consuming, especially when the number of buyers is large, we only collect the results of APX-OPT, when the number of buyers is no more than 400.

In the evaluations, we vary the number of buyers from 100 to 1000 with a step of 100, and uniformly distribute the buyers’ normalized valuations over $(0, 1]$. We also vary the proportion $p$ of the buyers’ tolerance bound to the number of their competitors, namely tolerance bound ratio, from 0 to 1 with a step of 0.1. All the results are averaged over 200 runs.

We consider three metrics, including social welfare, satisfaction ratio, and revenue. Social welfare is the sum of winning buyers’ valuations on the data. Satisfaction ratio is the percentage of winning buyers. Revenue is the sum of winning buyers’ payments.

5.1 Performance of DICTA-FUL

We evaluate DICTA-FUL using a complete competition graph with varying number of vertices/buyers. Figure 2(a) shows that the social welfare obtained by DICTA-FUL increases with the number of buyers and the tolerance bound ratio. This trend is natural since the increase of the number
of buyers and the tolerance bound both make more buyers obtain data.

Figure 2(b) demonstrates that the satisfaction ratio obtained by DICTA-FUL keeps almost stable with the varying number of buyers, but increases linearly with the tolerance bound ratio. We note that in full competition markets, the number of winners is \( p(n-1)+1 \). So, the satisfaction ratio is \( p(n-1)+1 \), which increases linearly with \( p \) and keeps almost stable with \( n \), especially when \( n \) is large.

Figure 2(c) illustrates that the revenue obtained by DICTA-FUL increases with the number of buyers. However, given the number of buyers, the revenue first increases when the tolerance bound ratio is no more than 0.5, and then gradually decreases to 0 when the tolerance bound ratio reaches 1. This is because when the tolerance bound ratio is relatively small, the growth of winning buyers does not significantly decrease the payments. However, after passing 0.5, the increment of tolerance bound ratio makes that the revenue generated by additional winners cannot even cover the revenue loss from existing winners due to dramatic drop of the payments.

5.2 Performance of DICTA-PAR

For this set of evaluations, we crawl real competition relations from Google Finance (https://www.google.com/finance). Specifically, we start from one company (denoted by central company), and record its related companies. Then, we record the related companies of each of the newly appeared companies. We repeat this process until the number of companies recorded exceeds 1000. We choose eight main business sectors, and pick a representative company as central company for each of them, i.e., Google for Internet services (1), Morgan Stanley for financial services (2), Sony for electrical equipment (3), Volkswagen for automobile (4), Walmart for retail (5), Royal Dutch Shell for oil and gas (6), Procter&Gamble for consumer goods (7), and Coca-Cola for beverage (8).

Figure 3(a) shows that the social welfare obtained by DICTA-PAR and APX-OPT both increase with the number of buyers and the tolerance bound ratio \( p \). Furthermore, we can observe that DICTA-PAR achieves a social welfare very close to APX-OPT.

Figure 3(b) demonstrates that the satisfaction ratio obtained by DICTA-PAR and APX-OPT both keep almost stable with the number of buyers, and increase with the tolerance bound ratio. The reason is similar to what we have explained in Section 5.1. Additionally, DICTA-PAR again achieves a satisfaction ratio very close to APX-OPT.

Figure 3(c) illustrates that the revenue obtained by DICTA-PAR increases with the number of buyers. However, with the increase of tolerance bound ratio, the revenue increases before reaching 0.6, and then gradually decreases to 0 when the tolerance bound reaches 1. The reason is similar
to what we have discussed in Section 5.1. The rough turning point changes from 0.5 to 0.6, because partial competition is considered in this evaluation instead of full competition.

Figure 4 shows the evaluation results on social welfare among different sectors. With the same number of buyers, the difference among sectors on social welfare is negligible, which means that the competitions in different business sectors have similar distributions, and DICTA-PAR exhibits pretty stable performance in different sectors.

6. RELATED WORKS

Data can be reproduced at negligible marginal cost, thus in unlimited supply just as digital goods. Digital goods auctions have been widely studied during the last several years. Goldberg et al. [23] studied the single round and sealed bid auction of digital goods in which items are in unlimited supply. Goldberg and Hartline showed that as for commodity in unlimited supply, no constant-competitive truthful auction is envy-free [22]. Mehta and Vazirani proved that in digital goods setting, for any randomized auction, which is truthful in expectation, there exists an equivalent randomized auction which randomizes over truthful deterministic auctions [43]. Balkan and Blum presented approximation algorithms to maximize seller’s revenue in an unlimited supply setting [8]. Furthermore, there have been some works trying to turn any randomized mechanism into an asymmetric deterministic one with approximately the same revenue [2, 11]. The online variant of digital auctions have also been extensively studied [10, 13, 14, 16, 33]. Attribute auctions, which include a wide variety of discriminatory pricing problems, have been studied in [9, 13]. However, these digital goods auctions never consider the competitions among buyers.

Recently, Moor et al. [44] designed a double auction to allocate answers (from data providers) to queries in the Web of Data. However, they did not consider the externalities in data trading. There also exist some loosely related works about data pricing, but they do not involve auction mechanism design. We categorize these works to be data based pricing, which defines prices of information goods, and query based pricing, which defines prices of queries. The data based pricing schemes considered in [38, 48, 51] are static, which means that prices are not dynamically updated. There are also some works in dynamic setting [34, 45]. The query based pricing models, which allow the seller to set prices for a few views and to calculate the price of a query, are studied in [7, 35, 36, 37, 39, 40, 41, 42].

In spectrum auction, the conflict-free buyers can be allocated the same channel simultaneously [4, 18, 27, 31, 50, 52, 53, 54], which is similar with that the two buyers with no competition relationship can be allocated data at the same time. However, in spectrum auction, if there exists interference between two buyers, these two requests absolutely cannot be both satisfied. Instead, in our model, a buyer and her competitor can be allocated data at the same time, if only the number of her winner competitors is no more than her tolerance bound. Another significant difference is that in spectrum auction, the conflict graph is common knowledge; however, in data auction, the competition relationship between buyers is private information, which makes a buyer's valuation for the set of data decreases if their competitor is also allocated. Mechanism design with externalities have been widely studied in both economics and computer science literature. In economics literature, Jehiel et al. studied the mixed externalities in the sale of nuclear weapons [29], in which countries prefer their allies rather than their enemies to win the auction. The mechanism design with allocative externalities is also studied in [30]. Jehiel and Moldovanu [28] considered both allocative and information externalities in mechanism design. Although these works presented very general and powerful models, none of them addressed computational issues in the mechanism design problem. The problem of mechanism design with positive externalities has been studied in [3, 5, 12, 15, 26]. In computer science literature, a well-studied scenario with negative externalities is the sponsored search auction, in which a company’s valuation for being allocated an ad slot decreases if their competitor is also allocated [1, 6, 19, 20, 21, 24, 32]. We note that these works cannot be conveniently extended to data auctions for several reasons. At first, the number of ad slots is limited but data can be supplied unlimitedly. Secondly, most of these works were based on a cascaded model to simulate the user’s behavior that they will visually scan the list of ads from the top to the bottom. However, this model has nothing to do with data auctions and cannot be applied in data auctions absolutely.

Among these works, the one particularly relevant to our work is that of Ghosh and Mahdian [20]. They studied the post-click competition for conversions in the sponsored search. They designed expressive GSP-like mechanisms for the simplest form that an advertiser’s private value depends on exclusivity, i.e., whether her ad is shown exclusively, or along with other ads. Their auctions take as input two-dimensional bids for exclusive and nonexclusive display. However, their GSP-like mechanisms are not strategy-proof. On the contrary, our auction achieves dominant strategy incentive compatibility, and we adopt a relatively more expressive value function. Also, in our model, the competition relationship is private information, and we need to motivate the buyer to reveal their true competitor set.

7. CONCLUSIONS

In this paper, we have modeled the data trading market as an auction with negative externalities. We have studied two different but connected market scenarios, including full competition markets and partial competition markets. For full competition markets, we have designed DICTA-FUL to compute the optimal allocation, and integrate it with the celebrated VCG mechanism. Thus, DICTA-FUL achieves both strategy-proofness and optimal social welfare. For partial competition markets, we have shown that finding the optimal allocation is NP-hard, and also hard to approximate. In this scenario, we have designed DICTA-PAR, which is a combination of a $d$-approximation allocation algorithm and a carefully designed charging scheme. Evaluation results based on both artificial and real competition relations have shown that both of the two proposed auction mechanisms achieve good performance, in terms of social welfare, satisfaction ratio, and revenue.
REFERENCES


[31] G. S. Kasbekar and S. Sarkar. Spectrum pricing games with bandwidth uncertainty and spatial reuse in...


