A Survey of Truthfulness in Differentially Private Mechanism Design

Annabel Li Edward Lue

***Abstract*—In recent years, differential privacy has become a well-accepted standard for the protection of sensitive information. As our capacity to collect and analyze data improves, rigorous privacy budgets and mechanisms to improve truthfulness become valuable topics of research. In this survey, we will examine how truthfulness and differential privacy are intrinsically connected in the field of mechanism design for several widely-studied game theory models—crowdsensing, auctions, and matching. This paper will introduce mechanisms with various truthfulness and privacy guarantees, evaluate the latest progress in theoretical and empirical development for different game applications, and present further avenues for research.**

I. Introduction

In the past, mechanism design largely ignored privacy as an incentive in a player's utility functions. However, in reality, people often factor the leakage of private information into their decision-making. Recent research has investigated many mechanisms and privacy definitions to rectify this missing piece in traditional mechanism design. This paper reviews truthful algorithms for private versions of three problems: crowdsensing, auctions, and matching. We characterize each problem and the development of truthful and private solutions in the context of private mechanism design as a whole. Finally, we discuss the current and future challenges and potential opportunities for research in the field.

Contributions.

1) We perform a comprehensive survey of truthful mechanisms in several privacy games. Unlike other surveys of private mechanism design, we specifically address guarantees of truthfulness in different private games. Since other surveys cover significantly broader topics, our survey is more comprehensive of truthful algorithms in private mechanism design.
2) We provide analysis of the past and present field of truthful and private mechanism design. Using this analysis, we provide a future outlook for truthful and private mechanisms, highlighting current challenges and unexplored problems.

II. Background

The theory of private algorithms has seen explosive growth in recent years. In particular, Dwork's [3] notion of *differential privacy* has proven to be an invaluable tool for measuring the privacy of users in a large number of applications:

Definition 1 ((ε, δ)-differential privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{H}$ is said to be (ε, δ)-differentially private ((ε, δ)-DP) if, whenever $D, D' \in \mathcal{D}$ are datasets differing on the data of only one user, we have:

$$\Pr[\mathcal{M}(D) \in H] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in H] + \delta$$

for all $H \subseteq \mathcal{H}$. If $\delta = 0$, we say $\mathcal{M}$ is ε-DP.

The following variant is also useful:

Definition 2 (LDP)**.** A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{H}$ is said to be (ε, δ)-locally DP ((ε, δ)-LDP) if, whenever $D, D' \in \mathcal{D}$ are *any pair* of datasets (considered to be the data of a single user), we have:

$$\Pr[\mathcal{M}(D) \in H] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in H] + \delta$$

for all $H \subseteq \mathcal{H}$. If $\delta = 0$, we say $\mathcal{M}$ is ε-LDP.

The following game-theoretical definition is also important to the results surveyed by this paper:

Definition 3 (η-dominant-strategy truthful)**.** Let T be the space of possible data from user input, and let t_{-i} be a vector of elements from T. Let $t_i, t_i' \in T$ be user i's actual and reported data respectively. Let u_i be user i's utility function. Then a mechanism $\mathcal{M} : T^n \to O$ is said to be η-dominant-strategy truthful (η-DST) if

$$u_i(\mathcal{M}(t_{-i}, t_i)) \geq u_i(\mathcal{M}(t_{-i}, t_i')) - \eta$$

regardless of the t_i' chosen. That is, a dishonest actor in an η-DST mechanism cannot gain more than η utility by lying.

III. Games and Associated Mechanisms

A. Crowdsensing

In the era of mobile devices, there is an incredible opportunity for gathering sensory data. However, location information associated with the sensory data is often sensitive. Thus, researchers have developed private mechanisms for the release of sensory and location information. Of course, guaranteeing truthfulness in this game is of utmost importance. Truthful reporting of location produces an optimal selection of sensor data for social utility.

There are several problems that must be solved to effectively implement private crowdsensing. First and foremost, there is the issue of preserving user privacy. Second, there is the problem of task assignment, where the server must match users with sensing tasks. Task assignment is not necessary in all crowdsensing paradigms. In passive paradigms, users simply collect data by leaving a sensor, often a smartphone,

on and passively collecting data. However, spatial paradigms require users to go to specific locations to record data, which requires assigning users to tasks. Finally, there is the problem of quality control. Different users may have varying qualities of data (due to hardware differences, environmental noise, etc.) and therefore should be aggregated together differently.

Singla and Krause [13] addressed privacy and truthfulness in a passive crowdsensing model. Unlike previous non-private mechanisms, their protocol did not require users to directly reveal private information unless they were selected by the algorithm for data acquisition. Selected users would be compensated for their information, making the leakage of information worthwhile. Before selection, users only send obfuscated, locally differentially private data to the server and a bid for the "price" of the data. The selection algorithm greedily selects users based on the marginal gain per cost given the users' bids and the server's value function for each piece of sensor data. The selection algorithm also computes a payment for each selected user which will always provide an expected gain in utility for the user relative to their bid. In addition, the paper shows that the algorithm is truthful, meaning that users cannot gain expected utility by reporting a bid that differs from their personal cost of leaking their private information. Of course, the added obfuscation of the algorithm incurs some loss of utility. The paper found that the loss of utility was bounded empirically by 25%.

Wang et al. [15] designed a truthful and private algorithm for the spatial crowdsensing paradigm. In the spatial crowdsensing paradigm, there is a set of tasks, each of which are located at different Cartesian coordinates. Users must travel to task locations, which incurs some cost that increases proportionally with distance from the task. In the described mechanism, instead of obfuscating user locations, users report distances, obfuscated with Laplace noise, to a subset of the tasks that they are willing to perform. The benefit of this choice is that it is more resistant to inference attacks. In addition, the algorithm allows users to select a personal privacy budget, which will affect the Laplace noise that is added, depending on their valuation of privacy. Given these inputs, the paper describes an algorithm which selects winners by assigning each user to at most one task. This portion of the algorithm, which is called the probabilistic winner selection mechanism (PWSM) in the paper, occurs in two steps. First, the probability that a given user is the closest to a given task is computed and users are greedily assigned to all tasks where they are the most likely to be closest. This creates the issue of users who are assigned to multiple tasks, which is not allowed in this model. To resolve this, the algorithm considers all possible assignments of the user to a single task for which they are the closest user and then assigns the second closest user for the other tasks. This process is repeated iteratively to generate all possible situations. Then, the algorithm compares the total expected distance between every pair of situations and selects the situation with the lowest total expected distance. The second portion of the algorithm, called the Vickrey Payment Determination Mechanism (VPDM), ascribes payment for the winners determined by PWSM. The payment is determined in two parts: a distance compensation and a privacy compensation. The privacy compensation is directly computed from the user's selected privacy budget. The distance compensation is computed to require that the user has some chance P of receiving positive utility. The sum of these two values guarantees truthfulness and positive expected utility for selected users.

Overall, both crowdsensing algorithms are similar in structure and guarantees. They differ in the crowdsensing paradigm in which they function. The spatial crowdsensing paradigm is a partial generalization of the passive crowdsensing paradigm because it considers the addition of user distances from task locations. The exact paradigms considered in the papers have other nuanced differences such as budget constraints. However, these differences are generally insignificant.

Another interesting method of recovering true values without actually guaranteeing truthfulness is to use ideas from quality control. Li et al. [8] designed an algorithm that uses a weighted aggregation function based on inferred quality of user data after Gaussian perturbation. Untruthful users will slowly have their weight decreased, which will allow the server to learn the true value for environmental data. We will not discuss the details of weight inference as it does not guarantee truth from users. However, it is worth mentioning as it is another method of deriving true values in a crowdsensing model given perturbed user data.

B. Auctions

The auction has been heavily studied in mechanism design, and many auction formats have been designed for different situations and priorities. This section discusses single-item, multi-item, and digital goods auctions, focusing on well-known truthful mechanisms for auctions and their intersections with differential privacy. In mechanism design, truthful mechanisms are also referred to as incentive-compatible, meaning that agents receive optimal payoff by bidding truthfully.

A similar problem to auctions is also discussed—the discrete facility location problem. This problem determines a selection of locations for one or more facilities such that the selection minimizes agents' distances to the nearest facility.

In a single-item auction, the true values that mechanism designers want to know are the maximum amounts that bidders are willing to pay for the auction item—their actual valuations for the item. As an example, a common truthful mechanism that ensures participants bid their actual valuation is the sealed-bid second-price auction, also called the Vickrey auction [14]. In a sealed-bid first-price auction, each participant bids a value in secret and the highest bidder pays the value they bid. In a sealed-bid second-price auction, the highest bidder only has to pay the second highest price. While the optimal strategy for a first-price auction might be to underbid, the strategy for a second-price auction is for bidders to bid their actual valuation, in order to maximize their chances of winning the item while avoiding paying more than they believe the item is worth.

The sealed-price second-price auction is a variant of the Vickrey-Clarke-Groves (VCG) auction, generalized by Clarke [2] and Groves [5], which is a type of multi-item auction utilizing the VCG truthful mechanism. The multi-item auction is one in which multiple items must be distributed to multiple people, where each person receives at most one item. The VCG mechanism aligns the self-interest of the bidders with overall social welfare by determining payments based on individual social cost. Thus, the VCG mechanism is powerful in many cases but fails when bidders have the ability to collude and lie in conjunction with each other. This is where differential privacy can play a role by restricting the damage of collusion [9].

Definition 4 (collusion resistance)**.** For any mechanism M giving ε-differential privacy and any non-negative function g of its range, for any D_1 and D_2 differing on at most t inputs,

$$\mathrm{E}[g(M(D_1))] \leq e^{\varepsilon t}\,\mathrm{E}[g(M(D_2))]$$

In the case, g is the sum of all individual utilities. Collusion between t individuals has a limited impact on collective utility.

McSherry and Talwar [9] developed the exponential mechanism, which can apply to several types of auctions. For instance, in a single-item auction, the winner is chosen with exponential probability. They also note that the exponential mechanism can also be used in digital goods auctions, where the auctioneer has an unlimited amount of an appealing good and must set its price and choose recipients. Their results show that this mechanism, and any mechanism that is ε-differentially private, can result in $(\exp(\varepsilon) - 1)$-dominant-strategy truthfulness. Bidders have less incentive to lie because their influence on the outcome is bounded. However, this is not a guarantee of truthfulness, nor is collusion resistance a guarantee that collusion cannot occur. Rather, the benefits of lying and colluding are quantitatively bounded due to ε-differential privacy.

Nissim et al. [10] expand on McSherry and Talwar's research by combining their exponential mechanism with a commitment mechanism to achieve exact truthfulness. The commitment mechanism randomly chooses a decision for the auctioneer to make, and restricts the bidders into certain sub-optimal reactions when they are untruthful. For example, in the discrete facility problem, the agents could be forced to gain utility from their reported location, not their actual location [1]. They prove that, following this mechanism, telling the truth always has a greater utility than misreporting for both digital goods auctions and the facility location problem.

So far, the mechanisms by McSherry and Talwar [9] and Nissim et al. [10] do not achieve both exact truthfulness and ε-differential privacy. The next two mechanisms we will discuss successfully achieve both properties but can only apply to certain game models.

Xiao [16] describes a method to transform truthful mechanisms into truthful and ε-differentially private mechanisms. He describes his method for the discrete facility problem on a line, where each agent submits a location on the line and a third-party chooses a facility location to minimize combined distance to the facility. Each agent is trying to minimize their own distance and there is a finite number of locations for agents to choose from. The current truthful mechanism for this problem is for the third-party to choose the left-most median point; however, this is not ε-differentially private. Xiao suggests creating a histogram counting the number of players who choose each location and applying two-sided geometric noise to each bin. From the perturbed histogram, choosing the left-most median point would ensure both truthfulness and privacy.

Huang and Kannan [6] prove that the exponential mechanism can be both truthful and ε-differentially private when the objective function is maximizing social welfare. Social welfare is defined as the sum of every agent's valuations. This works in the same way that the VCG mechanism guarantees truthfulness—by making payments represent individual social cost. First, the outcome, the allocation decision, is chosen with probability proportional to the exponential of social welfare. Then the pricing for an agent is determined, in oversimplified terms, by subtracting everyone else's bids from everyone else's valuations so that the only way to maximize personal utility is by maximizing social welfare and bidding truthfully. The actual pricing scheme is more complex and involves the formula for Shannon entropy. Unexpectedly, their proof for exact truthfulness and ε-differential privacy from this mechanism involves a comparison to Gibbs free energy.

In comparing these mechanisms, we will examine their advantages and disadvantages for truthfulness and privacy bounds, possible implementations, and generalization. The four mechanisms can be divided into two categories. The first category includes the exponential mechanism by McSherry and Talwar [9], as well as the combination of the exponential and commitment mechanisms by Nissim et al [10]. These are more general algorithms that cannot guarantee both privacy and truthfulness. The second category includes the mechanism by Huang and Kannan [6] and the mechanism by Xiao [16]. Both achieve exact truthfulness and ε-differential privacy, but are difficult to generalize.

McSherry and Talwar's [9] exponential mechanism is mostly included in this evaluation as a basis for comparison. The exponential mechanism is a tool for differential privacy, and since it is ε-differentially private for $\varepsilon \leq 1$, it also happens to be ε-approximately dominant strategy truthful. This claim applies to all differentially private mechanisms. In response, Nissim et al. [10] point to the fact that a truthful strategy that is ε-approximately dominant means that misreporting is another dominant strategy. In a digital goods auction, the lower bound on revenue, which McSherry and Talwar find to be $OPT - 3\ln(e + \varepsilon^2 OPTn)/\varepsilon$ with a set of n bidders where OPT is the optimal price, is inferior to a mechanism with exact truthfulness. However, there are also limitations to the mechanism by Nissim et al. Due to the commitment mechanism, the choice of reaction by each agent no longer satisfies ε-differential privacy since their reaction space is restricted. Furthermore, their general mechanism cannot extend

to large type sets, except in specific cases, and cannot apply to cases where the objective function is sensitive, such as revenue maximization in single-item auctions.

When the algorithm by Xiao [16] adopts the VCG mechanism as its truthful mechanism, the results from Xiao and Huang and Kannan [6] become very similar. Both algorithms are private, truthful, and computationally-efficient. In game theory, another definition of efficiency is the algorithm's distance from optimal global utility. Using the VCG mechanism, Xiao's algorithm is (ε, δ)-differentially private and has an additive error on efficiency of $O(q \log(q/\delta)/\varepsilon)$, where q is the size of the type space. Huang and Kannan also show that their algorithm maximizes utility, which in their case is the social welfare function. The major limitation of Xiao's method is that it only works when the type space is small, and cannot be generalized to all mechanism design problems. In comparison, Huang and Kannan's method can only be used for social welfare maximization with payments and does not work well for multi-item auctions—it can only approximately implement the exponential mechanism and loses exact truthfulness and privacy.

C. Matching

Matching is yet another studied application of game-theoretic privacy. In a sense, these problems can be considered variants of the auction problems above: instead of matching participants with items that would grant them quantitative, measurable utility, participants are matched with schools or partners in a way that would grant them comparative utility. One class of matching problems, *stable* matching, deals with pair matching (usually one-to-one or one-to-many) that has the property that there should not exist pairs who mutually prefer each other over the partners that they are matched with. Again, it should be noted that in contrast to the previous section, both sides of the matching usually have comparative preferences, rather than one side having numeric utility associated with each option from the other side.

This class of problem was first discussed by Gale and Shapley [4], who considered the context of stable one-to-one bipartite preferential matching. Gale and Shapley [4] found that in bipartite one-to-one preferential matching, (a) there is always a stable matching and (b) such a matching can be found by a deferred-acceptance algorithm that is dominant-strategy-truthful for exactly one side. Roth [11] found that it is actually impossible to create a mechanism for creating stable matches in the one-to-one setting where the mechanism is simultaneously truthful for both sides. Even worse, Roth [12] found that in one-to-many matching it was impossible to create an algorithm optimal for the matchable-with-many side that incentivized truthful reporting for *either* side.

Kannan et al. [7] discuss the application of a slightly modified one-to-many stable-matching setting: they permit that participants on the side that may be matched with many (henceforth schools) may have some small number of unfilled, unstable spots to be occupied by participants from the side that may only be matched with one (henceforth students). Kannan et al. [7] then modify the definition of η-DST to be applicable here (where A are the n students and U are the m schools, with $\odot$ representing the "not attending school" option):

Definition 5 (η-approximately-dominant-strategy truthful)**.** Let $\mathcal{M}$ be a randomized algorithm mapping vectors of student and school ranked preferences $\succ$ to student-school matchings $\mu \in (U \cup \{\odot\})^n$. If we have for all students $a \in A$, all preference vectors $\succ$, all $[0,1]$-bounded utility functions $\nu_a : (U \cup \{\odot\}) \to [0,1]$ *consistent with* $\succ_a$ (that is, $\nu_a(u_1) \geq \nu_a(u_2)$ whenever $u_1 \succ_a u_2$), and any $\succ'_a \neq \succ_a$, that

$$\mathrm{E}_{\mu \sim \mathcal{M}(\succ)}[\nu_a(\mu(a))] \geq \mathrm{E}_{\mu \sim \mathcal{M}(\succ'_a, \succ_{-a})}[\nu_a(\mu(a))] - \eta,$$

then we say that $\mathcal{M}$ is η-approximately dominant strategy truthful (η-ADST).

It is easy to see that this is the comparative analogue of the η-DST definition under expectation. In some sense, it is stronger since it has to hold for **all** compatible utility functions; however, it again only has to do so under expectation. Remarkably, Kannan et al. [7] find a connection between ε-DP algorithms for selecting *admissions thresholds* for students (Kannan et al. consider school preferences to instead be the ranking induced by an admissions test where each student knows their own numeric score, so thresholds induce a matching where each student attends their favorite school that they "made the cut" for) and selecting η-ADST mechanisms for computing student-school assignments:

Theorem (Theorem 4.1 from [7])**.** Let $\mathcal{M}$ be any (ε, δ)-DP mechanism which takes in preference (with scores as above) vectors $\succ$ and outputs *admissions thresholds* from $\mathbb{R}^m_{\geq 0}$. Let $\mathcal{F}_\succ : \mathbb{R}^m_{\geq 0} \to (U \cup \{\odot\})^n$ be the function that computes the induced matching where each student attends their favorite school that they scored above the threshold for. Then the mechanism $\mathcal{F}_\succ \circ \mathcal{M}$ is $(\varepsilon + \delta)$-ADST.

This is the remarkably strong analogue of McSherry and Talwar's [9] connection in the comparative setting. Indeed, it achieves a linear trade-off in the privacy budget ε, rather than an exponential one. Kannan et al. [7] use this to exhibit several private, approximately truthful (in the students) stable matching algorithms that favor schools.

IV. Analysis and Conclusion

Research of all the problems discussed in this paper exhibit two different general trends. The first is the fact that research in crowdsensing is trending towards generalization of problems and their solutions. For instance, private and truthful crowdsensing has been generalized from the passive paradigm to the spatial paradigm introducing an additional cost for users traveling to task locations. The second is an opposite trend in auction, resource allocation, and matching models towards more complex mechanisms designed for stronger guarantees in specific cases. The combination of the exponential and VCG mechanisms resulted in stronger guarantees, but only for the case of social welfare maximization. In general, these problems use specificity to guarantee privacy, truthfulness, and

efficiency. Then, once there are algorithms that have strong guarantees, research generalizes the guarantees to broader problems. We anticipate these trends to continue for all these problems and for other privacy games.

Some areas that have not yet been researched include generalization of current problems with more complex utility functions, designing private and truthful algorithms that are compatible with multiple problems, and focusing mechanism design for specific game models. Another area we believe to be heavily under-researched is the optimization of existing algorithms. Differentially private algorithms inherently will come at the expense of utility when compared with an algorithm where data is freely shared with a third-party or server. Existing algorithms can be optimized to perform better empirically while maintaining privacy and truthfulness guarantees. Finally, even within these three problems, there is potential for cross-pollination of ideas. For instance, the core idea that privacy implies truthfulness by McSherry and Talwar was first designed in relation to digital goods auctions, but this weak bound also inspires a stronger result for one-to-many matching problems. Future research should consider repurposing ideas from one privacy game to another. As our world becomes increasingly data-driven, these future solutions to the challenges of data privacy and truthfulness become ever more critical.

REFERENCES

[1] H. Chan, A. Filos-Ratsikas, B. Li, M. Li, and C. Wang, "Mechanism design for facility location problems: A survey," in *arXiv preprint arXiv:2106.03457*, 2021.

[2] E. Clarke, "Multipart pricing of public goods," in *Public Choice*, 1971.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[4] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962. [Online]. Available: http://www.jstor.org/stable/2312726

[5] T. Groves, "Incentives in teams," in *Econometrica*, 1973.

[6] Z. Huang and S. Kannan, "The exponential mechanism for social welfare: Private, truthful, and nearly optimal," in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 2012.

[7] S. Kannan, J. Morgenstern, A. Roth, and Z. S. Wu, "Approximately stable, school optimal, and student-truthful many-to-one matchings (via differential privacy)," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '15. USA: Society for Industrial and Applied Mathematics, 2015, p. 1890–1903.

[8] Y. Li, H. Xiao, Z. Qin, C. Miao, L. Su, J. Gao, K. Ren, and B. Ding, "Towards differentially private truth discovery for crowd sensing systems," in *IEEE 40th International Conference on Distributed Computing Systems*, 2018.

[9] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science*, 2007.

[10] K. Nissim, R. Smorodinsky, and M. Tennenholtz, "Approximately optimal mechanism design via differential privacy," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012.

[11] A. E. Roth, "The economics of matching: Stability and incentives," *Mathematics of Operations Research*, vol. 7, no. 4, pp. 617–628, 1982. [Online]. Available: http://www.jstor.org/stable/3689483

[12] A. E. Roth, "The evolution of the labor market for medical interns and residents: A case study in game theory," *Journal of Political Economy*, vol. 92, no. 6, pp. 991–1016, 1984. [Online]. Available: https://doi.org/10.1086/261272

[13] A. Singla and A. Krause, "Incentives for privacy tradeoff in community sensing," in *IEEE Transactions on Mobile Computing*, 2013.

[14] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," in *The Journal of Finance*, 1961.

[15] Z. Wang, J. Hu, R. Lv, J. Wei, Q. Wang, D. Yang, and H. Qi, "Personalized privacy-preserving task allocation for mobile crowdsensing," in *IEEE Transactions on Mobile Computing*, 2019.

[16] D. Xiao, "Is privacy compatible with truthfulness?" in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, 2013.