web: *http://www.cis.upenn.edu/~aaroth* aaroth@cis.upenn.edu

Foundations for Private, Fair, and Robust Data Science

Aaron Roth

Much of modern machine learning and statistics is based on the following paradigm: the algorithm designer specifies an objective function, and then optimizes it over some class of models. This is a powerful methodology, but while it generally results in a tool that is exceedingly good as measured by the designers narrow objective function, when the optimization is performed over a rich class of models, the result can have unintended and unanticipated side effects. This is especially problematic when — as is increasingly the norm — the models are trained on people's sensitive data, or deployed to make important decisions about peoples lives. In these cases, the "side effects" can manifest themselves as gross violations of the social norms that we would expect of human beings occupying the same parts of the decision making pipelines that we are ceding to algorithms: norms like *privacy* and *fairness*. This training paradigm also assumes that the environment that the algorithm will operate on is static, and so can have unanticipated consequences when the resulting algorithms and models are deployed in dynamic environments, in which people change their behavior in response to the incentives engendered by the algorithm. And machine learning pipelines make it easy and tempting to re-use the same datasets over and over again, which can lead to spurious conclusions.

The main thrust of my research is to discover principled ways to avoid this kind of misbehavior. This includes embedding constraints of "privacy", "fairness", and other norms directly into the design of algorithms, using game theoretic reasoning to make predictions about the effects of algorithmic interventions in dynamic environments, and developing algorithmic principles that lead to rigorous statistical guarantees when data can be dynamically re-used. This broad research program involves at least three distinct exercises:

1. **Thinking Carefully about Definitions**: Words like "privacy" and "fairness" have imprecise and nuanced meanings, but constraints placed on algorithms must be mathematically precise. Much of my recent work — especially in algorithmic fairness — has focused on finding mathematical constraints that can represent a meaningful promise made to individuals, while still being statistically and computationally sound: i.e. we should be able to give algorithms that satisfy the definitions we adopt — not just on the data we have seen, but also out of sample.

2. **Deriving and Analyzing Algorithms**: Once we have a formalization of a particular kind of "fairness" or "privacy", a particular game theoretic model, or a particular notion of algorithmic robustness, we need to be able to perform useful computations that satisfy our desiderata. This is an algorithm design task, relying on tools from theoretical computer science, probability theory, and optimization. We often want to prove not only that our algorithms satisfy our definitions, but that they solve some problem *optimally* subject to their constraints.

3. **Mapping Out Tradeoffs**: A lesson one quickly learns in this area is that nothing comes for free. Definitions of fairness, privacy, and robustness are actually parameterized: one can ask for these things to varying quantitative degrees. But asking for a little more "fairness" or "privacy" will usually come at a cost: often to the accuracy of some analysis, but sometimes also to e.g. other incomparable notions of fairness. Part of my work involves mapping out optimal tradeoffs between different desiderata — both in the worst case, using mathematical tools, and on particular datasets using empirical methods.

Finally, one of the joys of being a professor is communicating discoveries to others and facilitating their adoption in practice. This involves writing for other researchers, who can join and enrich the technical community working on these problems, but also writing for the general public, and working with industry partners on technology transfer. I've had the opportunity to do all of these things. I've written two books so far. The first (*The Algorithmic Foundations of Differential Privacy* [DR14]) has become the authoritative technical reference for differential privacy. The second (*The Ethical Algorithm* [KR19]) is a general audience book that covers work in algorithmic fairness, privacy, game theory, and statistics, and advocates for a scientific approach towards embedding social norms as constraints directly into algorithms. I have also helped to bring differential privacy from theory to practice by collaborating with industry. I have advised Apple as they incorporated differential privacy into iOS10, and Facebook's Election Research Commission in collaboration with

Social Science One to share election-relevant data with researchers with the protections of differential privacy. I serve as the scientific advisor to several startups working to deploy differential privacy broadly. In the remainder of this research statement, I will summarize some of the highlights of my contributions to each of my four main areas of study.

Private Data Analysis: It has been difficult for medical research to reap the fruits of large-scale data science because the relevant "data" is often highly sensitive individual patient records which are restricted by law from being shared freely. Similar difficulties plague many other domains to which data science could be productively applied: for example, the US Census Bureau collects information about every American household, which is both tremendously useful for social science, as well as for informing the best distribution of resources — but is legally obligated to release only privacy preserving statistics about this valuable data. Facebook has an enormous social network dataset that could be leveraged to study some of the most important questions of the day — including the spread of "fake news" and the dynamics of election interference — but cannot make this data publicly available because of legitimate privacy concerns. What do we do? To leverage data science, do we need to give up on privacy?

The answer is no: Over the past 15 years, a statistical notion of privacy called "differential privacy" [DMNS06] has emerged and been developed, offering a mathematically precise, meaningful notion of privacy that is consistent with performing useful computations on data. I have been and continue to be a central contributor to this literature (I wrote the first PhD thesis on the topic [Rot10b], and continue to regularly publish and advise students in the area) — and more recently, in addition to continuing to push forward the basic theory, I have been actively involved in putting the techniques I helped develop into practice.

Informally, data analyses that are conducted subject to differential privacy promise individuals a strong form of plausible deniability. The promise of differential privacy is that no one observing the outcome of a computation—no matter their background knowledge—can determine substantially more accurately than random guessing whether a particular individual's data was included in the data analysis or not. It represents a promise of privacy by taking the position that computations that are performed independently of a particular individual's data cannot be said to violate their privacy; by extension, if there is no statistically significant difference between this hypothetical world in which a computation is perfectly private for that individual and the actual world in which the computation is performed on all user data, then it should not be thought of as substantially violating the privacy of any particular individual. Differential privacy provides a quantitative measure of how similar the ideal world — in which an individual's data is not used at all — and the real world are, statistically.

Remarkably, essentially any statistical data analysis task (informally, any task whose optimal solution depends only on the *data distribution*, and not on the particular individuals present in the data set) can be carried out subject to the protections of differential privacy, albeit at a cost. That cost usually manifests itself as a need for more data (given the same accuracy goal), and sometimes also as a need for more computational power. My research in the area focuses on foundational questions, with a particular focus on the following:

1. **Which problems can be solved subject to differential privacy, and how severe are the inevitable tradeoffs?** The initial reaction of many researchers to differential privacy was that it provided a strong guarantee of privacy, but that it was *too strong* to perform useful statistical calculations. My early work in the field aimed at showing that this was not the case, by giving surprisingly powerful algorithms for a broad set of basic private data analysis tasks. This has included the problems of synthetic data generation [BLR13, GHRU13, Rot10a] and interactive query answering [RR10, BR13, GRU12, HRU13, HR14], combinatorial [GLM$^+$10] and convex optimization [HHR$^+$16, HHRW16, HRRU14], low rank matrix approximation [HR12, HR13], and basic questions about how different differentially private sub-routines can be adaptively composed [RRUV16] and evaluated as a function of their output [LNR$^+$17].

2. **To what extent can computations be distributed and the basic trust model relaxed**? Most of the initial academic work on differential privacy studied the setting in which there was a central curator who was trusted to have direct access to the data. But the first large deployments of differential privacy — at both Google and Apple — did not conform to this setting. Instead they operated in the distributed ("local") model of differential privacy, in which there is no trusted curator, and randomization is applied on device. Another thrust of

my research has investigated the power of this model — first by doing some of the initial work on the query answering problem [GHRU13] and the "heavy hitters" problem in the local model [HKR12], which has become the canonical algorithmic task subsequently studied and deployed in the local model. More recently we have studied both practically motivated questions in this model, like the extent to which statistics can be persistently updated without expending additional privacy budget [JRUW18] and foundational questions on the power and necessity of interaction [JMNR19, JMR20] in the local model.

3. **Can we leverage the powerful suite of (non-private) optimization heuristics?** Many basic problems in private data analysis that are information theoretically feasible are computationally hard in the worst-case, including private synthetic data generation. This mirrors the state of the field in a number of other areas, including machine learning and operations research: fundamental problems are hard in the worst case. But in the case of machine learning and operations research, this has not impeded practical progress, because despite worst-case hardness, we have developed an extraordinarily powerful suite of optimization heuristics for problems like empirical risk minimization, SAT solving, and integer program solving. Can we leverage these powerful heuristics for hard problems in private data analysis? The challenge is that differential privacy is not something that can be established empirically, and must be proven in the worst case — even though the heuristics that we wish to leverage have no worst-case guarantees. We call algorithms that are computationally efficient granting access to a heuristic optimization oracle, but which are differentially private even in the worst case *oracle efficient*. In [GGH$^+$16], we developed the first oracle-efficient algorithm for the problem of synthetic data generation consistent with 3-way marginals. More recently, we substantially generalized this paradigm [NRW19], giving oracle efficient algorithms for generating synthetic data for every family of queries that has a combinatorial structure called a "universal identification set", which includes full d-way marginals, and a generic method for converting algorithms that are only private assuming the success of the heuristic to algorithms that are private in the worst case. We also extended a classical technique called "objective perturbation" to be able to solve *non-convex* ERM problems in an oracle efficient manner [NRVW19].

4. **How and when can the definition of differential privacy be relaxed in a principled manner?** Despite the remarkable success of differential privacy, there are settings in which the definition is either too strong to be compatible with certain kinds of tasks, or misparameterized to tightly express the desired privacy guarantees. For example, consider the problem of using private preference data to match people to the items they desire: this plainly cannot be solved (non-trivially) subject to differential privacy because the item desired by an individual is both the secret they wish to keep private, and the thing that must be included in a high quality matching. But problems like this exhibit special structure, in that not only the inputs, but also the outputs to the problem are distributed amongst the interested parties. In this setting, we defined *joint differential privacy* [KPRU14], which informally guarantees for each individual i, that they have differential privacy against the coalition of all *other* individuals $j \neq i$, so long as the message sent to individual i (e.g. the good *they* are matched to) remains secret. We have shown that matching and allocation problems [HHR$^+$16], separable convex programs [HHRW16], stable matching problems [KMRW15], equilibrium computation problems [KPRU14, RR14, CKRW15], and optimal flow and pricing problems [RRUW15] can all be solved subject to joint differential privacy, although they cannot be solved with the traditional notion of differential privacy. We have also shown that there are other natural problems (*exchange* problems) that cannot be solved subject to joint differential privacy, but *can* be solved under what we call *marginal* differential privacy — which requires only privacy from each individual $j \neq i$ in isolation, not coalitions of such individuals [KMRR18]. Beyond this family of relaxations, we have considered problems motivated by national security use-cases in which guarantees of privacy are only to be granted to most citizens (not, e.g. the people who might be the explicit target of an intelligence investigation) [KRWY16], in contrast to the standard guarantee which must be applied to all individuals equally. We have also considered a reparameterized family of differential-privacy like definitions, using the language of hypothesis testing, which (unlike differential privacy) is able to tightly track the guarantees of differential privacy under composition, and which has a number of other nice analytic properties [DRS19].

We have also developed a number of applications of differential privacy to other areas, including game theory and mechanism design, and robust data analysis, which we discuss in subsequent sections. Finally, I have had an ongoing and fruitful collaboration with Benjamin Pierce, Justin Hsu, Andreas Haeberlen, and colleagues developing programming languages able to certify the differential privacy (and related) properties of programs that can be expressed within them [BGA^{+}15, BGA^{+}16, WCHRP17, ZRH^{+}19].

Algorithmic Fairness: In the last decade, machine learning has made the transition from a tool generally used for making low stakes decisions (like spam filtering and targeted advertising) whose accuracy only matters in aggregate, to a tool used to make high stakes decisions about peoples lives, including in lending, hiring, and criminal sentencing. And anecdotes about deployed algorithms systematically exacerbating inequality are now commonplace: Julia Angwin's team at Propublica discovered that the COMPAS recidivism prediction tool used as part of inmate parole decisions in Broward County Florida had a substantially higher false positive rate on African Americans compared to the general population. An automated resume screening tool developed at Amazon was found to downweight resume's containing the word “women”, as in e.g. “Women's chess team”. The algorithms used to automatically assign credit limits used by the Apple Card are suspected of systematic gender bias. The examples are so abundant that we now *expect* unconstrained error minimization to result in models that systematically make a disproportionate number of their errors on some structured population of people.

So what should we do about it? Despite the volume and velocity of recent research on this topic, there is little agreement. The state of the art is roughly where the study of data privacy was 20 years ago, and in my view we have yet to find the “right” definitions. Because of this, my work in the area has been very much focused on definitions, and my main research thrust has been to try and find definitions that simultaniously offer semantic guarantees to individuals, while being *actionable* — i.e. algorithmically and statistically satisfiable without the need to make heroic and unjustified assumptions about the data.

The most popular family of fairness definitions are what I call “statistical” definitions of fairness. At a high level, the statistical approach to fairness follows this template: partition the world into a small number of protected groups (often broken down by race or gender), pick some statistic of a classifier (false positive and negative rates have become popular), and then ask that this statistic be approximately equalized across the protected groups. This is an attractive approach in large part because it is immediately actionable: verifying statistical fairness requires only estimating a small number of averages, and although their are interesting computational challenges involved in optimizing subject to these constraints, there are no fundamental obstacles. But this approach to fairness promises essentially nothing to individuals, because the constraints bind only over coarse averages. Consider the simple goal of (say) selecting individuals to approve for a loan, with the idea that to be “fair” by both gender (Male and Female) and color (Blue and Green) we should equalize the loan approval rate across both axes. A solution accomplishing this is pictured in Figure 1: but this is cold comfort to a green man, who will be denied a loan with certainty. This is a cartoon, but similar effects emerge in real data — what we call “Fairness Gerrymandering” [KNRW18].

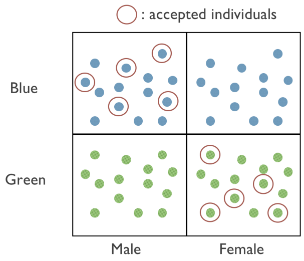

Figure 1: “Fair”?

A major thrust of my research in this area has been to come up with actionable *individual* notions of fairness. Individual fairness constraints — first proposed by [DHP^{+}12] — are constraints that bind on individuals, and hence provide semantic guarantees to particular people. But the initial proposal — informally that “similar individuals be treated similarly” has been difficult to realize in part because it pre-supposes the existence of a “task-specific similarity metric” that seems difficult to define.

In our first fairness paper [JKMR16] and our subsequent elaborations [KRW17, JJK^{+}17, JKM^{+}18] we propose that in many problems, there is already a notion of “merit” built into the model: namely, the unknown labels that the algorithm is trying to predict. This suggests a variant of the original constraint proposed by [DHP^{+}12], that informally translates into “no individual should be preferentially favored over any other except on the basis of true qualification”. We called this constraint “weakly meritocratic fairness”, and showed that it could be non-trivially achieved in bandit learning problems (both simple and contextual) [JKMR16, JKM^{+}18] and selection problems [KRW17], and that it could be extended to take into account long-term effects in the framework of reinforcement learning [JJK^{+}17]. A weakness of this approach, however, is that it takes the labels very seriously as a measure of “merit”, which is not appropriate in many situations, and often requires strong assumptions on the relationship between features and labels.

We then turned our attention to fairness constraints that could be achieved without any assumptions at all. A natural reaction to the example of the failure of statistical notions of fairness in Figure 1 is to suggest that we should simply have specified more groups to be protected. But which ones? If we specify all possible groups, no learning is possible without overfitting. It turns out that this statistical problem can be overcome by asking for statistical fairness over an infinite collection of groups, so long as that collection of groups is "simple" (has bounded VC-dimension). Optimizing subject to this constraint becomes non-trivial (indeed, it is not even clear how to efficiently *check* whether a fixed classifier satisfies this notion), but we were able to show that it is efficiently reducible to the problem of unconstrained learning by giving an oracle efficient algorithm [KNRW18] and conducting a series of experiments on real data [KNRW19].

The above proposal is practical and was well received, but in the end only mitigates, but does not eliminate, the "fairness gerrymandering" problem. This is because it does not ultimately abandon the statistical approach to fairness. More recently, we identified a way to offer truly individual-level guarantees without making assumptions about the data, in the special case in which individuals are subject not just to one, but to many classification tasks of similar stakes (think of targeted advertising as a key example). The constraint that we call "average individual fairness" informally asks that false positive rates (or any other popular measure of statistical fairness) be equalized — not across groups, but across individuals [KRSM19]. This is possible because in this setting we can redefine *rate* to refer to an expectation over a distribution of classification tasks, for a fixed individual. In this setting, we are also able to give oracle efficient algorithms, as well as novel generalization guarantees that hold not just over the distribution over individuals (as is standard), but also over the distribution of problems [KRSM19].

Another ongoing line of work attempts to satisfy the original notion of individual fairness due to [DHP$^+$12] in the absence of an explicitly specified "fairness metric". In [GJKR18], we assume that there is a human being who cannot explicitly enunciate a metric, but who can make judgements of similarity that are implicitly consistent with some metric. We show that under the assumption that the metric has a simple form, there is an efficient algorithm that obtains optimal regret in contextual bandit problems while satisfying the implicit fairness constraint. In [JKN$^+$19] we remove all of the assumptions from [GJKR18] about the form of the human beings' judgements (which no longer need to be consistent with any metric, let alone a simple one, and no longer even need to be consistent with one another, which allows for multiple human judges) and give an oracle efficient algorithm for optimally trading off classification accuracy with the frequency with which the algorithm makes decisions that violate the human beings' subjective notion of fairness.

I also have worked on (and have an ongoing interest in) different notions of fairness in settings that go beyond classification, which have been understudied in the literature. This includes applying statistical notions of fairness to online learning settings with censored feedback [BLR$^+$19], in allocation problems with finite supply [EJJ$^+$19], in the presence of additional constraints (like differential privacy) [JKM$^+$19], and "unfairness" that can arise because of the need to explore in bandit learning settings, and situations in which this tension can be avoided [KMR$^+$18]. Finally, in addition to some expository work [CR18, BHJ$^+$18, KR19], I have a line of work employing tools from game theory to understand fairness related problems — I discuss this in a subsequent section.

Robust Data Analysis: Large swaths of the social science literature are now in upheaval due to a replicability crises: a distressingly large number of published papers purporting to find statistically significant findings fail to replicate when data analyses are repeated on new data sets. This has in large part been blamed on "p-hacking" and "researcher degrees of freedom", which broadly refers to the freedom that researchers have in choosing which analyses to perform on what variables, often after examining the data itself. This kind of "adaptive" data analysis invalidates classical means of correcting for multiple hypothesis testing and guaranteeing statistical validity, informally because one must in principle correct for *all analyses that might have been performed had the results come out differently* — which is essentially impossible when human researchers are making decisions on the fly.

My collaborators and I discovered a remarkable connection between statistical validity in adaptive data analysis and differential privacy several years ago, in a paper that we published in *Science* [DFH$^+$15c]. In a nutshell, the theory we developed over a series of papers [DFH$^+$15c, DFH$^+$15b, DFH$^+$15a, DFH$^+$17] guarantees that analyses that were conducted with the protections of differential privacy, and that are accurate with respect to the data sample on hand, *are guaranteed also to be accurate on the data distribution from which the dataset was drawn*. In other words, differentially private analyses cannot overfit, and are immune from p-hacking. This was surprising at the time, but is natural in retrospect: the goal of private data analysis is to derive generalizable insights about the underlying data distribution, while explicitly avoiding being sensitive to the particulars of the dataset on hand: a goal perfectly aligned with generalizable data analysis. The power of this approach lies in the fact that differential privacy allows for a broad range of data analyses, which can be composed adaptively, allowing researchers to interactively explore data while avoiding the dangers of p-hacking.

Our initial paper and subsequent follow-up work provided an "in-principle" possibility result, in that it gave a theorem which was remarkable in its asymptotic performance — but one that was limited in the analyses it applied to, and didn't beat even naive baselines on realistic dataset sizes. My work since then has focused on developing the theory further to bring it closer to practice. This has included broadening the assumptions on the data analyses to which transfer theorems apply (beyond differential privacy) [CLN$^+$16], broadening the class of analyses for which differential privacy guarantees adaptive generalization (all the way to arbitrary hypothesis testing) [RRST16], generalizing the kinds of adaptivity that differential privacy provably protects against (from adaptive query selection to adaptive data gathering) [NR18], and giving empirical procedures that can guarantee provable data-dependent bounds on generalization performance that can be substantially tighter then the best possible worst-case bounds [RRS$^+$19].

Most recently, we came up with a new proof of the basic "transfer theorem" connecting differential privacy to generalization, which underlies what has come to be a small research area, which improves by several orders of magnitude on the concrete worst-case bounds, making them plausibly useful on reasonable dataset sizes [JLN$^+$20]. The new proof provides a fresh perspective through which to study the effects of differential privacy on data analysis, which I am optimistic will be useful for future work. Briefly, it observes that the dataset can always be viewed as being freshly sampled from the conditional data distribution, *conditioned* on the transcript of the analyses that have been performed so far, after analysis has concluded — and then shows that differential privacy guarantees that this conditional distribution is close to the original data distribution. This reverses the order of events (in actuality, first the data is sampled, and *then* analyses are conducted), putting them back into the standard regime of non-adaptive data analysis.

Game Theoretic Reasoning: The most basic assumption in traditional machine learning is that the environment in which an algorithm is trained is identical to the environment in which it will act — and in particular, that both the distribution on individual features, and the relationship between features and labels can be relied on to be the same in both training and deployment. This is the basis of the generalization guarantees that rationalize the entire endeavor — but is also plainly false in settings in which the data provided to the algorithm originates from human decision making. Human beings are rational and self interested agents, and change their behavior when doing so is beneficial to their own interests. These effects could often be safely ignored when deployed algorithms were making mostly obscure or inconsequential decisions that didn't merit much human attention — but they come to the fore as we start using algorithms to make decisions that are highly consequential to people. It is therefore important to be able to 1) predict the outcomes that will result from the interaction of large numbers of self interested agents in the presence of particular incentives, and 2) design algorithms that encode incentives that will result in the outcomes that we want.

My early work in game theory focused on question 1: what kinds of outcomes can we reliably predict when self interested agents interact, given that we expect them to be computationally bounded, and traditional notions of game theoretic equilibrium (like Nash equilibrium) can be hard to compute? Early on, we showed that even under the mild assumption that individuals acted according to learning dynamics that have "low regret", it is possible to make strong predictions about the welfare of the resulting outcome [BHLR08, Rot08], which could sometimes be refined under small deviations from rationality [CLPR08]. And although I have continued to have an interest in problems in equilibrium prediction and computation [RBKM10, GR16] and other basic problems in game theory and auction design [CLPR10, GRST10, DIR14, BBR15], the main thrust of my research in game theory has been at the interface of privacy, learning theory, and fairness.

Game Theory and Machine Learning: The standard assumption in learning theory is that feature/label pairs are drawn from a static distribution that is the same during training and deployment, and that is invariant to the classifier that the learner chooses. But the data that we train on is often generated by rational decision-makers in direct response to incentives. Consider, for example, the problem of predicting a customer's purchasing decisions given prices. If the prices are themselves drawn from a distribution, then this is not in principle different in kind from standard machine learning problems with a high dimensional label space. Yet it already introduces new complexities, because even when the customers have simple (e.g. linear) utility functions, the *labels* (i.e. bundles of purchased goods) we are trying to predict are complicated: they are the result of solving a constrained optimization problem. Beigman and Vohra [BV06] defined this problem and called it "learning from revealed preferences", and gave several information theoretic results. In [ZR12] we began the study of *efficiently* learning from revealed preferences, and gave the first computationally efficient learning algorithms for customers with linear utility functions. In [JRRW16], we gave algorithms for the substantially more general problem of learning solutions to arbitrary data-parameterized linear programs with unknown objective functions. Then, in [ACD$^+$15], we extended this work to the problem of dynamically setting prices in an online manner so as to maximize profit, when the only feedback to the algorithm remains the "revealed preferences" of the customers — i.e. the bundles that they purchase in response to prices. This now becomes a learning problem in which the data distribution changes in response to the choices of the algorithm. In [RUW16] and [RSUW17], we substantially generalized the class of utility functions (beyond linear) for which we could efficiently solve this high dimensional dynamic pricing problem, by giving a two-stage procedure that was able to employ tools from convex optimization and convex analysis, despite the non-convex nature of the problem. In [DRS$^+$18], we extended this "revealed preferences" approach to learning to the "strategic classification" problem that had been earlier introduced by [HMPW16]. The strategic classification problem studies a more standard binary classification problem, in which the individuals to be classified have preferences over the label they receive, and some ability to manipulate their features (think of spam filters as the canonical example). The goal is to deploy the classifier that maximizes accuracy *after* individuals adapt their behavior to the deployed classifier, or in other words, to compute an optimal *stackelberg equilibrium*. Prior work had studied this problem in the complete information setting — i.e. when the incentives and abilities of the agents to be classified were entirely known during the machine learning training process. We studied the problem in the revealed preferences setting, in which individuals with unknown incentives drawn from a fixed but unknown distribution must be classified in an online manner, and the only feedback received is their revealed preferences — i.e. their behavior in response to the deployed classifiers.

Game Theory and Privacy There are two ways in which the study of privacy interacts with game theory. The first is more obvious: the deployment of a privacy technology changes peoples incentives and costs for sharing data, and game theoretic reasoning is the right tool with which to capture these endogenous effects. That is, we can use game theory as a tool to study privacy. But it turns out that we can also use differential privacy as an algorithmic tool to solve game theoretic problems! Differential privacy is a measure of algorithmic robustness to *unilateral* changes in individual reports, which makes it cleanly map onto game theoretic notions of stability and equilibrium. I have worked extensively on both aspects of this connection, which Mallesh Pai and I surveyed in [PR13].

Game Theory for Privacy The differential privacy literature traditionally treats the dataset as a static object, already collected. But where does it come from, and how should the privacy parameter be chosen? A key property of differential privacy is that it lessens the individual costs for participation in a study, and therefore should make it easier to collect data. We can therefore view the process of collecting data — compensating people for their participation — and setting the privacy parameter together, as an economic problem. We originally formulated this problem as follows [GR13]: individuals own their own data points, and have unknown costs (parameterized by the privacy parameter ϵ) for allowing their data to be used in a differentially private computation. A data analyst wishes to *purchase* access to data so as to estimate some statistic of the population: she either wants to minimize the error of the statistic given a budget constraint, or wants to minimize cost given a fixed accuracy constraint. Note that here the data analyst has no explicit utility for privacy except insofar as it makes it cheaper to gather data. We gave optimal, truthful auctions solving both problems when the unknown individual costs for data usage did not themselves need to be kept private, and broad impossibility results when they did. Subsequently we extended this work in a number of ways [RS12, LR12, GLRS14, CLR$^+$15], and also considered more centralized methods to use economic reasoning (attempting to maximize social welfare) to select a privacy parameter [HGH$^+$14].

One weakness of the proceeding line of work is that we did not model *why* people had costs associated with the use of their data, we simply assumed in the model that they did. In [CLPR16] we studied a two-stage game in which individuals had no intrinsic value for privacy of their actions in the first round, except insofar as it effected their equilibrium utility in the second round. In this model, we were able to study the equilibrium effects of introducing a differentially private channel to information flow between the first and second round, and modifying the privacy parameter. We found counterintuitively that sometimes *increasing* the level of differential privacy could actually result in more information being revealed about players, and lower utility — because of the ways in which the stronger privacy protections influence individual behavior.

Privacy for Game Theory More surprisingly, as was first observed by McSherry and Talwar [MT07], differential privacy can be used as a tool to solve purely game theoretic problems, in which privacy is not an explicit goal. This is because differential privacy offers stability in the face of *unilateral deviations* by individual agents, which is the same kind of stability that equilibrium concepts like dominant strategy and Nash equilibria are based on. The original observation by [MT07] was used to design approximately dominant strategy truthful auctions, using the fact that differentially private selection rules automatically imply approximate dominant strategy truthfulness, because *misreporting* ones true valuation function can only result in a small change in the distribution over the selected outcome, and therefore at most a small increase in expected utility. But a substantial drawback of this kind of application of differential privacy is that it makes *everything* (not just truthful reporting) an approximate dominant strategy equilibrium, because it just de-couples the selected outcome from each individual's report. Our introduction of *joint differential privacy* in [KPRU14] addresses this problem: under joint differential privacy, the outcome relevant to agent i can depend arbitrarily on agent i's own report, but must be differentially private in the reports of others. We showed how to use jointly differentially private mechanisms in combination with traditional techniques from mechanism design to yield algorithms which make truthful reporting an approximately dominant strategy, while not making everything else an approximately dominant strategy as a side effect. We have shown how to solve a number of interesting game theoretic problems using the tools of joint differential privacy including equilibrium selection [KPRU14, RR14, CKRW15], truthful and optimal congestion pricing [RRUW15], school-optimal and approximately student truthful stable matchings [KMRW15] (this is provably impossible if one does not allow any relaxation to truthfulness), and truthful item pricings in combinatorial auctions [HHR$^+$16, HHRW16]. We also showed how imposing differential privacy on the signal structure of a repeated game causes the set of equilibria of the repeated game to collapse to the set of approximate equilibria of the one-shot game, which can dramatically improve the price of anarchy [PRU16].

Game Theory and Fairness: The majority of work in "algorithmic fairness" has studied algorithms as closed systems, under the assumptions that:

1. The party that desires "fairness" either is the algorithm designer herself, or else has complete control over the algorithm designer, and so can impose fairness constraints by fiat, and

2. That a change in the algorithm has no downstream or upstream effects on the decision-making pipeline within which the algorithm is embedded.

But these are idealized assumptions that are rarely true in practice. Figuring out how to robustly reason about realistic situations without these simplifying assumptions is difficult, and likely a decades-long research agenda. However, we have begun to make some early inroads. In [KKM$^+$17], we consider the case in which "fairness" is desired by a government agency who does not have the power to compel action by a third party (say, a lender), but does have the power to provide subsidies. In a model in which the government agency has only limited knowledge of the data available to the third party, we show how to provide minimum cost subsidies to a purely profit-minded agent so as to guarantee "weakly meritocratic fairness" at all but a bounded number of rounds.

In [KRZ19], we study how one might encourage various fairness desiderata while taking into account the "downstream" effects of the deployed algorithm. In our model, we imagine that we are in control of the admissions and grading policies of a school. However, there is a downstream employer who makes purely (Bayesian) rational decisions, *taking into account the policies we fix*. We show both positive and negative results in this model about the extent to which it is possible to enforce fairness constraints on the entire employment pipeline, informally representing either the constraint that "similarly talented students before the school admissions decision should have the same probability of making it all the way through the pipeline and obtaining a job at the end, regardless of which group they come from", and "the employer should have no incentive to make decisions as a function of group membership".

Finally, in [JKL+20], we study the "upstream" effects of deployed algorithms — i.e. how they can result in changes in human behavior that in turn effects the data that the algorithm is acting on. We take criminal justice as a motivating application, because this is a domain in which there has been much debate about the "right" notion of fairness we might want algorithms to satisfy. We study the question in a game-theoretic model, in which individual decisions about whether or not to commit crimes (and hence overall crime rates in different groups) are the rational response to the incentives set up by the criminal justice system and the outside opportunities available to the individual. We model different populations as differing only in their access to outside opportunities — which is enough to result in the differing base crime rates in different populations that are known to make various debated notions of fairness incomparable with one another [KMR17]. We then consider what the optimal policy would be (absent any explicit fairness desiderata) if the goal of the policy designer was to *minimize* the overall crime rate, across all groups. What we find is that (in contrast to policies optimized for predictive accuracy in static models), the optimal policy intentionally avoids taking group information into account, and ends up satisfying the popular notion of fairness that asks for equalized false positive and negative rates across groups — and pointedly does *not* equalize positive predictive value across groups, which is an alternative fairness notion that cannot be simultaneously guaranteed.

Going Forward: In the near term, I am excited to continue to pursue basic research in differential privacy, algorithmic fairness, game theory, and learning theory, and am excited about the prospects of expanding the applications of differential privacy to other areas. Many of the fundamental questions in data privacy (especially the extent to which it interacts with computational efficiency) remain unanswered, and we still only have a preliminary understanding of how markets for private data should function – and little understanding of how *privacy* behaves in equilibrium, when users change their behavior as a function of the mechanism proposed. Similarly, algorithmic work in "fairness" is still in a nescient stage — having yet to find the right definitions. Moreover, studying the upstream and downstream effects of fairness-motivated interventions has hardly begun, and I view this as an important research direction for the next several years. In the longer term, I am interested in exploring connections between differential privacy and other fields: algorithmic stability is an important phenomenon in many areas, and I believe it will continue to have important applications to statistics and machine learning. Much more work needs to be done to apply the connection we have discovered to core statistical questions. In the past decade, the differential privacy literature has developed a sophisticated toolkit for reasoning about a strong algorithmic stability constraint, and I expect that this toolkit will continue find much broader use.

References

[ACD+15] Kareem Amin, Rachel Cummings, Lili Dworkin, Michael Kearns, and Aaron Roth. Online learning and profit maximization from revealed preferences. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 770–776. AAAI Press, 2015.

[BBR15] Moshe Babaioff, Liad Blumrosen, and Aaron Roth. Auctions with online supply. *Games and Economic Behavior*, 90:227–246, 2015.

[BGA+15] Gilles Barthe, Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Pierre-Yves Strub. Higher-order approximate relational refinement types for mechanism design and differential privacy. In Sriram K. Rajamani and David Walker, editors, *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2015, Mumbai, India, January 15-17, 2015*, pages 55–68. ACM, 2015.

[BGA+16] Gilles Barthe, Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Pierre-Yves Strub. Computer-aided verification for mechanism design. In Yang

Cai and Adrian Vetta, editors, *Web and Internet Economics - 12th International Conference, WINE 2016, Montreal, Canada, December 11-14, 2016, Proceedings*, volume 10123 of *Lecture Notes in Computer Science*, pages 279–293. Springer, 2016.

[BHJ+18] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[BHLR08] A. Blum, M.T. Hajiaghayi, K. Ligett, and A. Roth. Regret minimization and the price of total anarchy. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 373–382. ACM New York, NY, USA, 2008.

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

[BLR+19] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, 2019.

[BR13] A. Blum and A. Roth. Fast private data release algorithms for sparse queries. *RANDOM*, 2013.

[BV06] Eyal Beigman and Rakesh Vohra. Learning from revealed preference. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 36–42. ACM, 2006.

[CKRW15] Rachel Cummings, Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Privacy and truthful equilibrium selection for aggregative games. In *International Conference on Web and Internet Economics*, pages 286–299. Springer, 2015.

[CLN+16] Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Proceedings of the 29th Conference on Learning Theory, COLT*, 2016.

[CLPR08] C. Chung, K. Ligett, K. Pruhs, and A. Roth. The Price of Stochastic Anarchy. In *Proceedings of the 1st International Symposium on Algorithmic Game Theory*, page 314. Springer-Verlag, 2008.

[CLPR10] C. Chung, K. Ligett, K. Pruhs, and A. Roth. The Power of Fair Pricing Mechanisms. In *Proceedings of the 9th Latin American Theoretical Informatics Symposium*, 2010.

[CLPR16] Rachel Cummings, Katrina Ligett, Mallesh M Pai, and Aaron Roth. The strange case of privacy in equilibrium models. In *Proceedings of the seventeenth ACM conference on Economics and computation*. ACM, 2016.

[CLR+15] Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani. Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 317–324. ACM, 2015.

[CR18] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[DFH+15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*. ACM, 2015.

[DFH+15c] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[DFH$^+$17] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The magazine archive includes every article published in communications of the acm for over the past 50 years. *Communications of the ACM*, 60(4):86–93, 2017.

[DHP$^+$12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[DIR14] Shaddin Dughmi, Nicole Immorlica, and Aaron Roth. Constrained signaling in auction design. pages 1341–1357, 2014.

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference TCC*, volume 3876 of *Lecture Notes in Computer Science*, page 265. Springer, 2006.

[DR14] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc, 2014.

[DRS$^+$18] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70. ACM, 2018.

[DRS19] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[EJJ$^+$19] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 170–179. ACM, 2019.

[GGH$^+$16] Marco Gaboardi, Emilio Jesús Gallego, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. Dual query: Practical private query release for high dimensional data. *Journal of Privacy and Confidentiality*, 7(2), 2016.

[GHRU13] A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately Releasing Conjunctions and the Statistical Query Barrier. *SIAM Journal on Computing (SICOMP)*, 2013.

[GJKR18] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

[GLM$^+$10] A. Gupta, K. Ligett, F. McSherry, A. Roth, and K. Talwar. Differentially Private Approximation Algorithms. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2010.

[GLRS14] Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. Buying private data without verification. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 931–948. ACM, 2014.

[GR13] Arpita Ghosh and Aaron Roth. Selling privacy at auction. *Games and Economic Behavior (GEB)*, 2013.

[GR16] Paul W. Goldberg and Aaron Roth. Bounds for the query complexity of approximate equilibria. *ACM Trans. Economics and Comput.*, 4(4):24, 2016.

[GRST10] Anupam Gupta, Aaron Roth, Grant Schoenebeck, and Kunal Talwar. Constrained non-monotone submodular maximization: Offline and secretary algorithms. In *International Workshop on Internet and Network Economics*, pages 246–257. Springer, 2010.

[GRU12] Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release. In *TCC*, pages 339–356, 2012.

[HGH+14] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE, 2014.

[HHR+16] Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

[HHRW16] Justin Hsu, Zhiyi Huang, Aaron Roth, and Zhiwei Steven Wu. Jointly private convex programming. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 580–599. SIAM, 2016.

[HKR12] Justin Hsu, Sanjeev Khanna, and Aaron Roth. Distributed private heavy hitters. In *39th International Collloquim on Automata, Languages, and Programming*, pages 461–472, 2012.

[HMPW16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122. ACM, 2016.

[HR12] Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *The 44th ACM Symposium on the Theory of Computing*, pages 1255–1268, 2012.

[HR13] Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *The 45th ACM Symposium on the Theory of Computing*, 2013.

[HR14] Zhiyi Huang and Aaron Roth. Exploiting metric structure for efficient private query release. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 523–534. Society for Industrial and Applied Mathematics, 2014.

[HRRU14] Justin Hsu, Aaron Roth, Tim Roughgarden, and Jonathan Ullman. Privately solving linear programs. In *41st International Collloquim on Automata, Languages, and Programming*, 2014.

[HRU13] Justin Hsu, Aaron Roth, and Jonathan Ullman. Differential privacy for the analyst via private equilibrium computation. In *The 45th ACM Symposium on the Theory of Computing*, 2013.

[JJK+17] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1617–1626. JMLR. org, 2017.

[JKL+20] Christopher Jung, Sampath Kannan, Changwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. *Manuscript*, 2020.

[JKM+18] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163. ACM, 2018.

[JKM+19] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008, 2019.

[JKMR16] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 2016.

[JKN+19] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

[JLN+20] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacyâĂŹs generalization guarantees. In *11th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 151, page 31, 2020.

[JMNR19] Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *Foundations of Computer Science (FOCS)*, 2019.

[JMR20] Matthew Joseph, Jieming Mao, and Aaron Roth. Exponential separations in local differential privacy through communication complexity. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2020.

[JRRW16] Shahin Jabbari, Ryan M Rogers, Aaron Roth, and Steven Z Wu. Learning from rational behavior: Predicting solutions to unknown linear programs. In *Advances in Neural Information Processing Systems*, pages 1570–1578, 2016.

[JRUW18] Matthew Joseph, Aaron Roth, Jonathan Ullman, and Bo Waggoner. Local differential privacy for evolving data. In *Advances in Neural Information Processing Systems*, pages 2375–2384, 2018.

[KKM+17] Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 369–386. ACM, 2017.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 43. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

[KMR+18] Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, pages 2227–2236, 2018.

[KMRR18] Sampath Kannan, Jamie Morgenstern, Ryan Rogers, and Aaron Roth. Private pareto optimal exchange. *ACM Transactions on Economics and Computation (TEAC)*, 6(3-4):12, 2018.

[KMRW15] Sampath Kannan, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven 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*, pages 1890–1903. Society for Industrial and Applied Mathematics, 2015.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2018.

[KNRW19] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109. ACM, 2019.

[KPRU14] Michael Kearns, Mallesh M Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. *American Economic Review*, 104(5), 2014.

[KR19] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, USA, 2019.

[KRSM19] Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems*, pages 8240–8249, 2019.

[KRW17] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1828–1836. JMLR. org, 2017.

[KRWY16] Michael Kearns, Aaron Roth, Zhiwei Steven Wu, and Grigory Yaroslavtsev. Private algorithms for the protected in social network search. *Proceedings of the National Academy of Sciences*, 113(4):913–918, 2016.

[KRZ19] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248. ACM, 2019.

[LNR$^+$17] Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Steven Z Wu. Accuracy first: Selecting a differential privacy level for accuracy constrained erm. In *Advances in Neural Information Processing Systems*, pages 2566–2576, 2017.

[LR12] Katrina Ligett and Aaron Roth. Take it or leave it: Running a survey when privacy comes at a cost. In *WINE*, pages 378–391, 2012.

[MT07] F. McSherry and K. Talwar. Mechanism Design via Differential Privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103, 2007.

[NR18] Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. In *International Conference on Machine Learning*, pages 3717–3726, 2018.

[NRVW19] Seth Neel, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. Differentially private objective perturbation: Beyond smoothness and convexity. *arXiv preprint arXiv:1909.01783*, 2019.

[NRW19] Seth Neel, Aaron Roth, and Zhiwei Steven Wu. How to use heuristics for differential privacy. In *Foundations of Computer Science (FOCS)*, 2019.

[PR13] Mallesh Pai and Aaron Roth. Privacy and mechanism design. *SIGecom Exchanges*, 2013.

[PRU16] Mallesh M. Pai, Aaron Roth, and Jonathan Ullman. An antifolk theorem for large repeated games. *ACM Trans. Econ. Comput.*, 5(2), October 2016.

[RBKM10] A. Roth, M.F. Balcan, A. Kalai, and Y. Mansour. On the Equilibria of Alternating Move Games. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2010.

[Rot08] Aaron Roth. The Price of Malice in Linear Congestion Games. In *Proceedings of the 4th International Workshop on Internet and Network Economics*, page 125. Springer-Verlag, 2008.

[Rot10a] A. Roth. Differential Privacy and the Fat Shattering Dimension of Linear Queries. In *Proceedings of the fourteenth annual workshop on randomization and computation (RANDOM 2010)*, 2010.

[Rot10b] Aaron Roth. *New algorithms for preserving differential privacy*. PhD thesis, Carnegie Mellon University, 2010.

[RR10] A. Roth and T. Roughgarden. Interactive Privacy via the Median Mechanism. In *The 42nd ACM Symposium on the Theory of Computing.*, 2010.

[RR14] Ryan M. Rogers and Aaron Roth. Asymptotically truthful equilibrium selection in large congestion games. In Moshe Babaioff, Vincent Conitzer, and David Easley, editors, *ACM Conference on Economics and Computation, EC '14, Stanford , CA, USA, June 8-12, 2014*, pages 771–782. ACM, 2014.

[RRS$^+$19] Ryan Rogers, Aaron Roth, Adam Smith, Nathan Srebro, Om Thakkar, and Blake Woodworth. Guaranteed validity for empirical approaches to adaptive data analysis. *arXiv preprint arXiv:1906.09231*, 2019.

[RRST16] Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494. IEEE, 2016.

[RRUV16] Ryan Rogers, Aaron Roth, Jonathan Ullman, and Salil Vadhan. Privacy odometers and filters: Pay-as-you-go composition. In *Advances in Neural Information Processing Systems*, 2016.

[RRUW15] Ryan Rogers, Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Inducing approximately optimal flow using truthful mediators. In *Proceedings of the sixteenth ACM conference on Economics and computation*. ACM, 2015.

[RS12] Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *ACM Conference on Electronic Commerce*, pages 826–843, 2012.

[RSUW17] Aaron Roth, Aleksandrs Slivkins, Jonathan Ullman, and Zhiwei Steven Wu. Multidimensional dynamic pricing for welfare maximization. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 519–536. ACM, 2017.

[RUW16] Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Watch and learn: optimizing from revealed preferences feedback. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 949–962. ACM, 2016.

[WCHRP17] Daniel Winograd-Cort, Andreas Haeberlen, Aaron Roth, and Benjamin C Pierce. A framework for adaptive differential privacy. *Proceedings of the ACM on Programming Languages*, 1(ICFP):10, 2017.

[ZR12] Morteza Zadimoghaddam and Aaron Roth. Efficiently learning from revealed preference. In *Internet and Network Economics*, pages 114–127. Springer, 2012.

[ZRH$^+$19] Hengchu Zhang, Edo Roth, Andreas Haeberlen, Benjamin C. Pierce, and Aaron Roth. Fuzzi: a three-level logic for differential privacy. *PACMPL*, 3(ICFP):93:1–93:28, 2019.