This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

Competitive Statistical Estimation with Strategic Data Sources

Tyler Westenbroek, *Student Member, IEEE*, Roy Dong,
Lillian J. Ratliff, *Member, IEEE*, and S. Shankar Sastry, *Fellow, IEEE*

***Abstract*—In recent years, data has played an increasingly important role in the economy as a good in its own right. In many settings, data aggregators cannot directly verify the quality of the data they purchase, nor the effort exerted by data sources when creating the data. Recent work has explored mechanisms to ensure that the data sources share high quality data with a single data aggregator, addressing the issue of moral hazard. Oftentimes, there is a unique, socially efficient solution.**

In this paper, we consider data markets where there is more than one data aggregator. Since data can be cheaply reproduced and transmitted once created, data sources may share the same data with more than one aggregator, leading to free-riding between data aggregators. This coupling can lead to non-uniqueness of equilibria and social inefficiency. We examine a particular class of mechanisms that have received study recently in the literature, and we characterize all the generalized Nash equilibria of the resulting data market. We show that, in contrast to the single-aggregator case, there is either infinitely many generalized Nash equilibria or none. We also provide necessary and sufficient conditions for all equilibria to be socially inefficient. In our analysis, we identify the components of these mechanisms which give rise to these undesirable outcomes, showing the need for research into mechanisms for competitive settings with multiple data purchasers and sellers.

I. Introduction

DATA has increasingly seen a role in the economy as an important good. As an input to machine learning algorithms, data can not only create new products and innovations, but also be used to redesign business strategies and processes. As the demand for data increases, we have seen the formation of data aggregators, who collate data for either use or resale. A fundamental information asymmetry arises between data aggregators and data sources: how can aggregators verify the quality of the data they purchase from data sources?

In particular, data sources often incur an effort cost to obtain high quality data. For example, devices require maintenance and upkeep to ensure accurate measurements, portable sensors need to use their limited energy resources to collect and transmit data, and human agents may need to be compensated to properly perform a desired task. As such, if a data aggregator wants a high quality data point, they must appropriately compensate the data source. Furthermore, this problem is complicated by the fact that the data aggregators cannot observe the effort exerted, and only the data received. As such, the payments must be calculated from the data sets alone, with no knowledge of the effort exerted or noise levels of data points. This problem has led to the design of a variety of mechanisms to ensure data sources provide quality data, which we will outline in more detail in Section II.

The contribution of this paper is the study of the data market that forms when multiple data aggregators share the same pool of data sources. In particular, we note that data is non-rivalrous, in the sense that it can be cheaply copied and shared with multiple data aggregators. Since a data aggregator does not 'consume' the good after purchasing it, data sources will have an incentive to share the same data with as many aggregators as are willing to pay. We show that the non-rivalrous nature of data introduces a coupling between data buyers: when a data aggregator incentivizes a data source to produce high quality data, other data aggregators benefit. In particular, this coupling leads to undesirable properties of the equilibrium. In many single-aggregator formulations, equilibria are unique and there is no social inefficiency. In contrast, the multiple-aggregator case leads to a multiplicity of equilibria, and social inefficiencies across all equilibria.

The rest of this paper is organized as follows. In Section II, we discuss the related literature and contextualize our contributions. In Section III, we introduce our model for data sources, data aggregators, and their interactions in the data market. In Section IV, we characterize the generalized Nash equilibria in the data market, and identify necessary and sufficient conditions for social ineffiency. In Section V, we extend the results to cases where data sources do not share their data with all data aggregators. Finally, we close with final remarks in Section VI.

II. Related Literature

In recent years, there has been a quickly growing body of literature on models for data exchange and data markets. Broadly speaking, the existing literature can be broken down by two categories: models with a single data purchaser and single data source, and models with a single data purchaser and multiple data sources.

In the first category, find a class of models which study a single data purchaser and a single data source. These works focus on the game theoretic interactions and information states between the two agents. In particular, these works consider the strategies arising from direct signals, actions, and payments, rather than indirect coupling that can arise from multiple sources or purchasers. Some of these papers feature multiple data sources, but these are ultimately separable into a collection of single-source models, and, at their core, focus

T. Westenbroek and S. S. Sastry are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, 94707, USA, e-mail: {westenbroekt,sastry}@eecs.berkeley.edu.

R. Dong is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, 61820, USA, e-mail: roydong@illinois.edu.

L. J. Ratliff is with the Department of Electrical Engineering, University of Washington, Seattle, WA, 98195, USA, e-mail: ratliffl@uw.edu.

This work is partially funded by NSF CNS:1656873.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

on the direct interactions between buyers and sellers of data. In [1], optimal mechanisms for a single data source to sell to a single buyer are developed using a signaling framework. The authors of [2] design a menu of prices for different data qualities, employing a screening framework. In [3], the authors consider a single aggregator and single source, and show how repeated interactions with noisy verification allow for mechanisms which elicit costly effort from a data source. A single data source charging data purchasers for queries about customer preferences is studied in [4].

In the second category, there are a class of models which study a single data purchaser with multiple data sources. These works focus on capturing how the data supplied by one data source affects another. In [5], the authors consider a single data aggregator and multiple data sources, and show how robustness of the sample median provides protection against strategic data sources. In [6], the authors consider a single data aggregator and multiple data sources in a setting with verifiable data, and allow the data and the cost of revealing data to be arbitrarily correlated.

There is also a new body of work in the single-aggregator, multiple-source case, using *peer prediction* mechanisms, first introduced in [7]. These techniques often use scoring techniques to evaluate the 'goodness' of received data, and often examine classification tasks. In [8], [9], the authors develop mechanisms for eliciting the truth in crowdsourcing applications, while [10]–[12] consider theoretical extensions to strengthen the original results of [7], all in the context of a single aggregator. In [13], the authors consider a classification problem with a single aggregator and multiple data sources, which extends the classic peer prediction results by exploiting correlations between the queries and query responses.

A parallel literature considers similar ideas in the regression domain. These works design general payment mechanisms, by which a central data aggregator may incentivize data sources to exert the effort necessary to produce and report readings which are deemed to be of high quality, with respect to the estimation task the aggregator is performing. The roots of these approaches can be traced least as far back as VCG mechanisms, a set of seminal results in mechanism design [14]. Indeed, numerous approaches for deciding payments based on the actions of other agents have been proposed [15]. Here, we again see attention given to crowdsourcing [16].

Several recent papers [3], [17]–[21] investigate new directions in this domain. In cases where, without the ability to directly determine the effort exerted by data sources, data buyers must design incentive mechanisms based solely on the data available to them. In [17], whose approach we extend here, the authors develop a mechanism which a data aggregator can use to precisely set the level of effort a collection of data sources exert when producing data. A similar mechanism is explored in [18]. Extensions are considered wherein data sources form coalitions [19], or where aggregators assess the quality of readings using a trusted data source [20]. Meanwhile, [21] and [3] investigate dynamic settings where data sources are repeatedly queried.

Our work is closest in spirit to the literature studying regression problems with multiple data sources, with our key contribution being the presence of multiple data aggregators that are coupled in their costs and actions. To our knowledge, this is one of the first papers which considers multiple data aggregators and multiple data sources simultaneously. In particular, we simultaneously model coupling between data aggregators in their cost functions, coupling in the payments to the same pool of data sources, and coupling between data sources due to payments that depend on their peers' data.

We suppose all data aggregators are trying to estimate the same function and share the same pool of data sources. Additionally, we assume each data aggregator has already chosen an estimator, and now must determine how to issue payments to have low estimation error with their exogenously fixed estimator. Our model builds heavily on the model introduced in [17], which featured a single data aggregator. Our contribution is an extension that models cases with multiple data aggregators. For consistency, we will refer to data purchasers as *data aggregators*, and data sellers as *data sources*.

Furthermore, the work in the paper is a significant extension of our prior work [22] where we considered strategic data sources with a specific exponential function mapping effort to query response quality. In the present work, we characterize equilibria and the price of anarchy for a much broader class of games between data buyers where the data sources' effort functions can be any non-negative, strictly decreasing, convex, and twice continuously differentiable function. The characterization we provide considers both bounded and unbounded feasible effort sets for the data sources.

III. Data Market Preliminaries

In this section, we outline the models for data sources, data aggregators, and the strategic interactions between them.

At a high level, each data aggregator collects data from data sources to construct an estimate of a given function. In exchange for this data, the data aggregator issues incentives to the data sources. The data aggregators have three terms in their cost function: 1) an estimation error term, which rewards the data aggregator for constructing a better estimate; 2) a competition term, which penalizes when other data aggregators have higher quality estimates; 3) a payment term, which is the cost incurred issuing incentives.

Each data source is able to produce a noisy sample of the desired function. The data sources can exert effort to reduce the variance of the data sample, and we assume the data sources are *effort-averse*, i.e. data sources will prefer to exert less effort, unless they are provided incentive by the aggregators. As such, the data sources have two terms in their utility function: 1) an incentive term, which rewards payments received; 2) an effort term, which penalizes effort exerted.

The level of effort exerted and the variance of the data are not known by the data aggregator; this *private information* gives rise to *moral hazard*. One of the problems for the aggregator is the task of designing incentives which depend only on the information available to them. Another important nuance is that data is *non-rivalrous*; thus, when a data source produces a higher-quality data sample, all the aggregators which receive this data benefit.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

In order to simplify the initial introduction of our model, we will first assume that each data source provides data to all the aggregators in the data market, and receives payment from all aggregators as well. In Section V, we will outline how our results change when this assumption is removed.

A. Overview

More formally, let $\mathcal{S} = \{1, \dots, N\}$ be the index set of *strategic data sources*, and let $\mathcal{B} = \{1, \dots, M\}$ be the index set of *strategic data aggregators*. Each data aggregator desires to construct an estimate for a given function $f \colon \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is a *feature space*. Practically, one may think of $\mathcal{D}$ as a set of features the data aggregators are capable of observing, while the mapping f encapsulates the relationship between the observable features and the outcome of interest.

Each data source $s \in \mathcal{S}$ is able to produce a noisy sample y_s of f at the fixed point $x_s \in \mathcal{D}$. The point x_s is common knowledge among all data sources and aggregators. The variance of y_s is proportional to the effort exerted by data source s to produce the reading. Each data source s is characterized by an *effort-to-variance* function $\sigma_s^2 : \mathcal{E}_s \to \mathbb{R}_{\geq 0}$, where $\mathcal{E}_s$ represents the set of feasible efforts that data source s can exert. When data source s exerts effort $e_s \in \mathcal{E}_s$, they produce the data point:

$$y_s(e_s) = f(x_s) + \epsilon_s(e_s) \tag{1}$$

Here, $\epsilon_s(e_s)$ is a random variable with mean 0 and variance $\sigma_s^2(e_s)$. The function σ_s^2 is common knowledge among all data sources and aggregators. However, while the function σ_s^2 is known, the effort exerted e_s is private. This means that the actual variance of y_s, namely $\sigma_s^2(e_s)$, is also private information of s. We will delve into assumptions in the data source model in greater detail in Section III-B.

Now, suppose a data aggregator is granted access to a data set $\{(x_s, y_s)\}_{s \in S}$. At this point, the data aggregator $b \in \mathcal{B}$ processes this data to construct an estimate for f. In exchange for this data set, the data aggregator issues payment $p_s^b(y)$ to data source s for each $s \in \mathcal{S}$. Here, $y = (y_1, \dots, y_N)$ denotes the data given to each member of $\mathcal{B}$. Note that the payment to s from b depends not only on the data supplied by s, but rather depends on all data available to b.

The data aggregator then incurs loss $L^b(p^b, e)$, which will depend on $p^b = (p_i^b)_{i \in \mathcal{S}}$, the payments issued, as well as $e = (e_i)_{i \in \mathcal{S}}$, the effort exerted by the data sources. We will formalize the data aggregator in greater detail in Section III-C.

The interaction of the data market proceeds in three stages.

1) *Aggregators declare incentives:* Each data aggregator $b \in \mathcal{B}$ commits to a payment contract $p^b = (p_i^b)_{i \in \mathcal{S}}$. The payments will depend on the data y shared with b, as well as the common knowledge information $x = (x_i)_{i \in \mathcal{S}}$ and functions $\sigma^2 = (\sigma_i^2)_{i \in \mathcal{S}}$.
2) *Sources exert effort, realize and share data:* In response to p^b, each data source s chooses an effort $e_s \in \mathcal{E}_s$. Then, the random variable y_s is realized according to (1). The data y_s is shared with each data aggregator. Note that s has control over y_s only through e_s. In other words, the data source chooses the quality of data they generate, but cannot arbitrarily manipulate the reported value of y_s.
3) *Aggregators construct estimates, issue payments:* Each data buyer b constructs their estimate $\hat{f}^b$, issues payments p^b to the data sources, and incurs loss L^b.

For convenience, we include a table summarizing the notation throughout this paper in Table I.

B. Strategic Data Sources

As mentioned previously, each data source $s \in \mathcal{S}$ has their own *feature vector* $x_s \in \mathcal{D}$, and samples the function f at this point. We may also refer to x_s as a *query* throughout the text, and y_s as the *query response* for data source s. The data source s is characterized by the *effort-to-variance* function $\sigma_s^2 : \mathcal{E}_s \to \mathbb{R}_{\geq 0}$. We assume $0 \in \mathcal{E}_s$ so that each data source may exert no effort in producing her reading if she desires.

Assumption 1. *For each $s \in \mathcal{S}$, the set $\mathcal{E}_s \subset \mathbb{R}_{\geq 0}$ is a closed, connected set and contains 0.*

Assumption 1 means that we consider two cases:

(i) $\mathcal{E}_s = [0, \infty)$, i.e. the data sources maximum allowed effort is unbounded.
(ii) $\mathcal{E}_s = [0, e_s^{\max}]$ for some $0 < e_s^{\max} < \infty$, i.e. the data sources maximum allowed effort is bounded.

Imposing an upper-bound on the amount of a effort a data source can exert can be used to model constraints such as hardware limitations. As we shall see in Section IV, the imposition of such constraints can drastically affect equilibrium behavior in the data market.

Once the data source s exerts effort $e_s \in \mathcal{E}_s$, they produce the data point y_s according to (1). Again, we note that the data source only controls the effort level e_s. They can only indirectly control y_s through e_s, and cannot report arbitrary values as their data. We also impose the assumption that the noise in the data is independent across data sources.

Assumption 2. *For each $s \in \mathcal{S}$, $\epsilon_s(e_s)$ is a random variable with mean 0 and variance $\sigma_s^2(e_s)$. Furthermore, the random variables $\{\epsilon_s(e_s)\}_{s \in \mathcal{S}}$ are independent.*

Both x_s and the function function σ_s^2 are common knowledge, but the effort e_s and $\sigma_s^2(e_s)$, the actual variance of y_s, are private.

For convenience, we let $\mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_N$ be the joint effort set and let $\sigma^2 = (\sigma_1^2, \dots, \sigma_N^2)$ be the tuple of effort-to-variance functions. We make the following assumptions on the effort-to-variance mappings σ^2.

Assumption 3. *For each data source $s \in \mathcal{S}$, the mapping $\sigma_s : \mathcal{E}_s \to \mathbb{R}_{\geq 0}$, which is the square root of σ_s^2, is (i) strictly decreasing, (ii) convex, and (iii) twice continuously differentiable.*

The assumptions correspond to the variance of the estimate generated by data source s decreasing in the effort exerted, with decreasing marginal returns.

Using the notation $p_s = (p_s^j)_{j \in \mathcal{B}}$, we model each data source with the following utility function:

$$u_s(e_s, p_s) = \mathbb{E}\Big(\textstyle\sum_{j \in \mathcal{B}} p_s^j(y(e)) \Big) - e_s \tag{2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

where the expectation is with respect to the randomness in y, the data generated by the data sources upon exerting effort e.[1] Note the form of (2) implies that the data sources are risk-neutral and effort-adverse. Additionally, the form of (2) also implies the effort e_s can be normalized to be comparable to the payments. We note that the timing of the game implies that data sources must commit to an effort level ex-ante.

Thus, in the second stage of the game, data source s has knowledge of the payment contracts $(p^b)_{b\in\mathcal{B}}$, and chooses e_s to maximize their $u_s(e_s, p_s)$, defined by (2). However, since the utility of each data source depends on the effort exerted by the other data sources, the payments $(p^b)_{b\in\mathcal{B}}$ induce a game between the data sources. In Section III-F we will fully characterize this game for the particular class of incentives we introduce in Section III-D.

C. Strategic Data Aggregators

The primary objective of each aggregator is to construct a low-variance estimate for the function f. We adopt the following formal definition for an estimator.

Definition 1 (Estimator [17])**.** *Let $\mathcal{H}$ be a family of functions $f : \mathcal{D} \to \mathbb{R}$. An estimator for $\mathcal{H}$ takes as input a collection $\mathcal{X} = (x_i, y_i)_{i=1}^N$ of examples $(x_i, y_i) \in \mathcal{D} \times \mathbb{R}$ and produces an estimated function $\hat{f}_{\mathcal{X}} \in \mathcal{H}$.*

As an example, $\mathcal{H}$ may be the class of linear functions $f : \mathbb{R}^n \to \mathbb{R}$, in which case one may produce an estimated function $\hat{f}_{\mathcal{X}} \in \mathcal{H}$ of f via linear regression.

Each data aggregator $b \in \mathcal{B}$ constructs his estimate for f from the class of functions $\mathcal{H}_b$, using the readings $\mathcal{X} = (x_s, y_s)_{s\in\mathcal{S}}$. We let $\hat{f}^b_{\mathcal{X}} \in \mathcal{H}_b$ denote the estimate that aggregator b constructs based on the readings they receive.[2]

Each data aggregator's estimator is given, fixed, and common knowledge among all agents. In other words, this means that, for each data aggregator, the process by which a data set is turned into an estimate is exogenous. We focus on the design of incentives once each buyer has chosen an estimator.

First, we introduce some restrictions on the class of estimators allowed. The following assumption is required for us to be able to consider the contribution of data source s to reducing aggregator b's estimation cost. Also, note that the functions $\{h_b\}_{b\in\mathcal{B}}$ will be non-negative by construction.

Assumption 4. *We assume the estimator for each $b \in \mathcal{B}$ is separable, in the following sense [17]. There exists a function h_b such that for all queries $\mathbf{x}$, distributions F over $\mathcal{D}$, and variances σ^2 of the reported estimates $\mathbf{y}$ at queries $\mathbf{x} = (x_i)_{i=1}^k$ in the dataset $\mathcal{X} = (\mathbf{x}, \mathbf{y})$:*

$$\mathbb{E}\left[\left(\hat{f}^b_{\mathcal{X}}(x^*) - f(x^*)\right)^2\right] = \sum_{i=1}^k h_b(x_i, \mathbf{x}, F)\sigma_i^2 \tag{3}$$

Here, the expectation is taken across the randomness in $\mathcal{X}$, as well as across $x^* \sim F$.

For brevity, we will also define the function g_b as follows:

$$g_b(\mathbf{x}, F, \sigma^2) = \textstyle\sum_{i=1}^k h_b(x_i, \mathbf{x}, F)\sigma_i^2 \tag{4}$$

Let $-b = \mathcal{B} \setminus \{b\}$ denote the index set of aggregators excluding b and let $p^{-b} = (p^j)_{j\in-b}$ be the payments of all aggregators excluding b. Aggregator b constructs payments so as to minimize:

$$\begin{aligned} L^b(p^b, e) = \mathbb{E}\Big[&\left(\hat{f}^b_{\mathcal{X}}(x^*) - f(x^*)\right)^2 \\ &- \textstyle\sum_{j\in-b} \zeta_j^b \left(\hat{f}^j_{\mathcal{X}}(x^*) - f(x^*)\right)^2 \\ &+ \eta^b \textstyle\sum_{s\in\mathcal{S}} p_s^b(y(e))\Big] \end{aligned} \tag{5}$$

As in (3), the expectation in (5) is taken with respect to $x^* \sim F_b$ and the randomness in the query responses y. The distribution F_b weighs the importance data aggregator b places on accurately estimating f for different query points $x \in \mathcal{D}$.

The scalars $\zeta_j^b \in [0, 1]$ parameterize the level of competition between aggregators b and j. When $\zeta_j^b = 0$, aggregator b is indifferent to the success of j's estimation; b interacts with j entirely through the incentives issued to the data sources. We note that, even when $\zeta_j^b = 0$ for all j and b, we can still see degeneracies and social inefficiency arise, since data aggregators will still be coupled through the data sources.[3] The parameter $\eta^b > 0$ denotes a conversion between dollar amounts allocated by the payment functions and the utility generated by the quality of the various estimates that are constructed. We make the assumption that aggregator b has knowledge of what estimator every other data aggregator plans to use, as well as the weighting distributions.[4]

D. Structure of Payment Contracts

Throughout this paper, we will assume a particular form for the payment contracts the aggregators offer to the data sources. Similar to previous notation, we let $-s = \mathcal{S} \setminus \{s\}$. For a given $b \in \mathcal{B}$ and $s \in \mathcal{S}$ we assume that p_s^b is of the form:

$$p_s^b(y^b) = c_s^b - a_s^b \left(y_s^b - \hat{f}^b_{\mathcal{X}_{-s}}(x_s)\right)^2 \tag{6}$$

Here, a_s^b and c_s^b are nonnegative scalars. Also, $\mathcal{X}_{-s} = (x_{-s}, y_{-s})$ denotes b's data set excluding s. Namely x_{-s} is the data features for all sources excluding s and $y_{-s} = (y_i)_{i\in-s}$ is the query responses to aggregator b, excluding s.

[1] For simplicity and as a first-step analysis, we assume that the data sources only care about the payments received from the aggregator, and are indifferent to which aggregators they share their data with. An interesting and practical extension would be to consider the case where the data sources' utility functions are aggregator-dependent. This could arise when data sources trust different aggregators differently, or over privacy concerns.

[2] In general, aggregators need not fit models of the same type—e.g., one data aggregator may choose to generate their estimate via linear regression, while another fits a polynomial of higher degree. Different estimator types across data aggregators may be used to encapsulate competitive advantages one has over another.

[3] This is a stylized formulation of how competition can affect different data aggregators, but we see interesting results arise even in this simple model. In the future, we hope to consider more extensive models of competition for data aggregators.

[4] This is a fairly strong assumption given that competing data aggregators are unlikely to inform their competitors how they intend to process the data supplied by the sources. Our work isolates how coupling between aggregators through data sources affect the data market; an interesting avenue for future work is to consider extensions with different information sets, and characterize the existence and severity of market inefficiencies in these various situations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

Note that these payments do not directly depend on the level of effort that any of the data sources exert, since the data aggregators do not have a means to directly observe these values. Rather, the payment to source s from aggregator b depends on the b's best estimate for $f(x_s)$ excluding s's data, namely, $\hat{f}^b_{\mathcal{X}_{-s}}(x_s)$. The payments only depend on the data reported to them, and can be calculated by the aggregator.

Similar payment contracts are common in the literature [17], [18], [20], in part because of their intuitive structure. The aggregator constructs an unbiased estimate of what data source s *should* report, and this estimate is not influenced by the data of s. This estimate is used to overcome the problem of moral hazard: all data sources are appropriately incentivized to reduce the variance of their reported data accordingly.

Given this payment structure, each data aggregator's choice of payment contracts reduces to choosing parameters (c^b, a^b) where $c^b = (c^b_i)_{i\in\mathcal{S}} \in \mathbb{R}^N$ and $a^b = (a^b_i)_{i\in\mathcal{S}} \in \mathbb{R}^N$.

In the single aggregator case (when $M = |\mathcal{B}| = 1$), it was shown in [17] that payments of the form in (6) induce a game between the data sources for which there is a unique dominant strategy equilibrium. That is, for each collection of parameters $(c^b_i)_{i\in\mathcal{S}}$ and $(a^b_i)_{i\in\mathcal{S}}$, the data sources each exert a unique level of effort. The authors develop and algorithm by which the single aggregator may select these parameters such that (i) data sources are incentivized to exert any level of effort that the aggregator desires, and (ii) data sources are compensated at exactly the value of their effort, i.e. $\mathbb{E}[p_s(y(e))] = e_s$.

This paper's contribution is the study of how pricing schemes of this form perform in the more general case where there is more than one data aggregator (when $M = |\mathcal{B}| > 1$), and data aggregators may compete with each other. The goal is to model multiple aggregators as strategic decision-makers in competition, and understand the *data market* where these agents interact. Thus, while prior work captured moral hazard, we extend this model to capture competition and the non-rivalrous nature of data.

E. Formulation of Aggregator Optimization Problem

As mentioned previously, the aggregators hope to minimize their costs, as given in (5). They do so by choosing the parameters (c^b, a^b). In this section, we will describe the aggregator's optimization problem in more detail, and specify constraints that the parameter choice must satisfy.

The first constraint is *individual rationality (IR)*. Individual rationality requires that each data source's utility is non-negative ex-ante [24].[5] This ensures that rational data sources are willing to exert effort to produce the data. The second constraint is non-negative payments from each data aggregator. Given that there are multiple aggregators, we introduce a constraint that the payment each aggregator offers to each s is non-negative ex-ante.[6]

[5]Alternatively, a data source's utility may be compared to an outside option; for simplicity, we model the outside option as having zero utility.

[6]Negative payments could be handled via exchangeable utilities among the data aggregators or via a trusted third–party to manage the allocations; however, in an effort to ensure clarity, we leave these scenarios aside.

We'll introduce some notation for brevity here; we let p^b_s denote the expected value of the payment p^b_s:

$$\begin{aligned}\mathrm{p}^b_s((c^b_s, a^b_s), e) &= \mathbb{E}[p^b_s(y)] \\ &= c^b_s - a^b_s\big(\sigma^2_s(e_s) + g_b(x_{-s}, \delta_{x_s}, \sigma^2_{-s}(e_{-s}))\big) \end{aligned} \tag{7}$$

where δ_x denotes the probability measure with mass one at x and $e = (e_i)_{i\in\mathcal{S}}$. Similar to previous conventions, we define:

$$\mathrm{p}_s((c_s, a_s), e) = \textstyle\sum_{b\in\mathcal{B}} \mathrm{p}^b_s((c^b_s, a^b_s), e)$$

Thus, the IR constraint for each data source s is formalized:

$$\mathrm{p}_s((c_s, a_s), e) \geq e_s \tag{8}$$

Similarly, the non-negativity constraint for each data source s and data aggregator b is given by:

$$\mathrm{p}^b_s((c_s, a_s), e) \geq 0 \tag{9}$$

The third constraint is *incentive compatibility (IC)*. Intuitively, IC states that when a data source is acting rationally and choosing actions to maximize their utility, they behave as the data aggregators intended. When there is a single aggregator, IC is typically enforced by the aggregator finding the effort that minimizes their cost, e^*_s, and then designing p_s such that $e^*_s = \arg\max_{e_s\in\mathcal{E}_s} \mathrm{p}_s((c_s, a_s), e) - e_s$.[7]

In the competitive setting, IC for one aggregator is defined holding all other aggregators payments fixed. Each of the data aggregators make their choice of payment subject to the fact that data source s selects effort according to

$$\max_{e_s\in\mathcal{E}_s} \textstyle\sum_{j\in\mathcal{B}} \mathrm{p}^j_s((c^j_s, a^j_s), e) - e_s \tag{10}$$

Note that the payment each source receives depends on the efforts exerted by the other data sources. Thus, for each set of contracts offered by the aggregators, a game is induced between the data sources to determine how much effort they will exert. The aggregators compete by issuing incentives, which influences the equilibrium behavior of this game.

From the perspective of the data aggregators, the IC constraint states the desired effort level e^*_s must be a *dominant strategy* for data source s; that is, e^*_s is the utility-maximizing action for s regardless of the actions taken by other sources $-s$. Formally, the following must hold for all $e_{-s} \in \mathcal{E}_{-s}$:

$$e^*_s = \arg\max_{e_s\in\mathcal{E}_s} \mathrm{p}_s((c_s, a_s), (e_s, e_{-s})) - e_s$$

With these constraints, we formulate a bilevel optimization problem for each aggregator. Consider a fixed aggregator $b \in \mathcal{B}$. Given a fixed action profile for all other buyers $-b$, i.e. given (c^{-b}, a^{-b}), aggregator b aims to solve:

$$\begin{aligned}\min_{(c^b, a^b)}\ & L^b((c^b, a^b), (c^{-b}, a^{-b})) \\ \text{s.t.}\quad & e^*_s = \arg\max_{e_s\in\mathcal{E}_s} \mathrm{p}_s((c_s, a_s), (e_s, e_{-s})) - e_s, \\ & \qquad \forall e_{-s} \in \mathcal{E}_{-s},\ \forall s \in \mathcal{S} \\ & \mathrm{p}_s((c_s, a_s), (e^*_s, e^*_{-s})) \geq e^*_s,\ \forall s \in \mathcal{S} \\ & \mathrm{p}^b_s((c_s, a_s), (e^*_s, e^*_{-s})) \geq 0,\ \forall s \in \mathcal{S} \\ & c^b_s \geq 0, a^b_s \geq 0,\ \forall s \in \mathcal{S}\end{aligned}$$

[7]For notational brevity, we will use $\arg\max$ as a function rather than a set-valued function throughout this paper; this is well-defined by Assumption 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

where L^b is defined in (5).

Note that this problem actually has N optimization problems as constraints, making is a difficult bilevel program. However, we will reformulate the aggregator's problem to a more manageable non-linear program in the sequel. This is possible, in part, due to the nice properties of the payment contract structure introduced in Section III-D; this tractability motivates the use of payment contracts of that particular form. Next, we analyze the induced game between the data sources and simplify the aggregator's optimization problem.

F. Induced Equilibrium Between Data Sources

To ensure a notion of *incentive compatibility in equilibrium*, we show there is a well-defined mapping from the parameters (c, a) chosen by the aggregators to the equilibrium e^*.

Definition 2. *For fixed payments $\{p_s^b\}_{s\in\mathcal{S},b\in\mathcal{B}}$, we say $e^* = (e_1^*, \ldots, e_N^*)$ is an* induced Nash equilibrium *if for each data source $s \in \mathcal{S}$:*

$$e_s^* = \arg\max_{e_s\in\mathcal{E}_s} \mathbb{E}\left[\sum_{j\in\mathcal{B}} p_s^j(y(e_s, e_{-s}^*))\right] - e_s \tag{11}$$

If (11) *holds for all $e_{-s} \in \mathcal{E}_{-s}$ rather than just at e_{-s}^*, then we say that e^* is an* induced dominant strategy equilibrium.

Suppose now that we have a set of payments of the form discussed in Section III-D, characterized by parameters (c, a). Data source s chooses effort e_s^* according to:

$$e_s^* = \operatorname*{argmax}_{e_s\in\mathcal{E}_s}\left[\sum_{b\in\mathcal{B}} c_s^b - a_s^b\big(\sigma_s^2(e_s) + g_b(x_{-s}, \delta_{x_s}, \sigma_{-s}^2(e_{-s}))\big)\right] - e_s \tag{12}$$

for each choice of $e_{-s} \in \mathcal{E}_{-s}$ made by the other data sources. It is straight forward to verify that (12) is a concave maximization problem which admits a unique globally optimal solution. This follows from our assumption that σ_s^2 is convex and decreasing, recalling that $a_s^b \geq 0$ for each $b \in \mathcal{B}$ and observing that $\mathcal{E}_s$ is a convex set. Moreover, note that the choice of this optimal effort e_s^* is not affected by the choice of e_{-s}, since each of the $g_b(x_{-s}, \delta_{x_s}, \sigma_{-s}^2(e_{-s}))$ terms enters (12) as a constant from the perspective of s. Thus, each choice of contract parameters selected by the aggregators leads to an induced dominant strategy equilibrium for the data sources. In particular, note that the choice of

$$\boldsymbol{a}_s = \textstyle\sum_{j\in\mathcal{B}} a_s^j \tag{13}$$

fully characterizes the level of effort that data source s exerts in equilibrium. We reiterate that the constraints on the aggregator's optimization problems will ensure the chosen contract parameters respect the IR and non-negativity constraints.

Next, we define $\mu_s \colon \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ to be the implicitly-defined map such that $\mu_s : \boldsymbol{a}_s \mapsto e_s^*$ returns the solution to (12) for a given choice of $\boldsymbol{a}_s \in \mathbb{R}_{>0}$. In the following section, we will use this mapping to simplify the optimization problem facing each of the aggregators.

Definition 3. *For a given data source $s \in \mathcal{S}$, let:*

$$\underline{\boldsymbol{a}}_s = \min\{\boldsymbol{a}_s \in \mathbb{R}_{>0} : \mu_s(\boldsymbol{a}_s) = 0\}. \tag{14}$$

When $\mathcal{E}_s = [0, e_s^{\max}]$ with $0 \leq e_s^{\max} < \infty$, define $\mathcal{A}_s = [\underline{\boldsymbol{a}}_s, \bar{\boldsymbol{a}}_s]$ where

$$\bar{\boldsymbol{a}}_s = \min\{\boldsymbol{a}_s \in \mathbb{R}_{>0} : \mu_s(\boldsymbol{a}_s) = e_s^{\max}\} \tag{15}$$

On the other hand, when $\mathcal{E}_s = \mathbb{R}_{\geq 0}$, define $\mathcal{A}_s = [\underline{\boldsymbol{a}}_s, \infty)$.

The above definition implies $\underline{\boldsymbol{a}}_s$ is the minimum value of $\boldsymbol{a}_s$ that the aggregators must offer data source s to ensure they do not have incentive to exert negative effort.[8] Similarly, if the aggregators increase $\boldsymbol{a}_s$ past $\bar{\boldsymbol{a}}_s$, source s cannot further increase the level of effort they exert, and the mapping μ_s ceases to be meaningful. Thus, when reformulating each buyers optimization in the following section we will additionally constrain $\boldsymbol{a}_s \in \mathcal{A}_s$ for each $s \in \mathcal{S}$.

The following lemma provides properties on the mapping μ_s which are needed to prove existence of equilibria for the game between aggregators in the first stage.

Lemma 1. *Fix a data source $s \in \mathcal{S}$. Then the mapping $\mu_s(\boldsymbol{a}_s)$ is continuous and strictly increasing in $\boldsymbol{a}_s$ for all values of $\boldsymbol{a}_s \in \mathcal{A}_s$.*

Proof. The first-order optimality condition for the data source is given by:

$$2\boldsymbol{a}_s\sigma_s(e_s)\tfrac{d}{de_s}\sigma_s(e_s) + 1 = 0 \tag{16}$$

By assumption σ_s is strictly decreasing and convex so that (16) has a unique solution for all $\boldsymbol{a}_s \in \mathcal{A}_s$. By definition, this solution is $\mu_s(\boldsymbol{a}_s)$. Implicit differentiation of (16) then yields:

$$\tfrac{d\mu_s}{d\boldsymbol{a}_s} = \left(2(\boldsymbol{a}_s)^2\left(\left(\tfrac{d}{d\mu_s}\sigma_s(\mu_s)\right)^2 + \sigma_s(\mu_s)\tfrac{d^2}{d^2\mu_s}\sigma_s(\mu_s)\right)\right)^{-1}$$

where we suppress the dependence of μ_s on $\boldsymbol{a}_s$. The right-hand side of the above equation is strictly positive by Assumption 3. Continuity follows directly by Assumption 3. □

G. Reformulation of Buyers Optimization Problem

Finally, using our previous analysis and assumptions, we reformulate the optimization problem faced by each aggregator. This reformulation will simplify our analysis of equilibrium behavior in the data market, and lend economic interpretability to the results presented in Section IV.

Previously, we assumed that aggregator b's estimator is separable in Assumption 4. This allows us to write the loss function of b as:

$$\begin{aligned} L^b((c^b, a^b), (c^{-b}, a^{-b})) &= \textstyle\sum_{i\in\mathcal{S}} h_b(x_i, x, F_b)\sigma_i^2(\mu_i(\boldsymbol{a}_i)) \\ &\quad - \textstyle\sum_{j\in-b}\zeta_j^b \sum_{i\in\mathcal{S}} h_j(x_i, x, F_j)\sigma_i^2(\mu_i(\boldsymbol{a}_i)) \\ &\quad + \eta^b\textstyle\sum_{i\in\mathcal{S}}\big(c_i^b - a_i^b\big[\sigma_i^2(\mu_i(\boldsymbol{a}_i)) \\ &\quad + \textstyle\sum_{l\in-i} h_b(x_l, x_{-i}, \delta_{x_i})\sigma_l^2(\mu_l(\boldsymbol{a}_l))\big]\big) \end{aligned}$$

Recall that $x = (x_i)_{i\in\mathcal{S}}$ is fixed and common knowledge. Thus, we can replace each of the evaluations of the h_j's with constants. Towards this end, for each $i, l \in \mathcal{S}$ and $b \in \mathcal{B}$, we define:

$$\beta_i^b = h_b(x_i, x, F_b) \tag{17}$$

[8]This situation could correspond to source s obfuscating their data, for example. We have restricted $\mathcal{E}$ to the non-negative orthant, so we will add constraints to ensure we are operating within the domain of our model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

$$\xi^b_{i,l} = \begin{cases} h_b(x_l, x_{-i}, \delta_{x_i}) & i \neq l \\ 1 & i = l \end{cases} \tag{18}$$

Note that each $\xi^j_{i,l} \geq 0$, by definition of the $\{h_b\}$. In addition, for each $i \in \mathbb{S}$ and $b \in \mathcal{B}$, define:

$$\gamma^b_i = \beta^b_i - \textstyle\sum_{j \in -b} \zeta^b_j \beta^j_i \tag{19}$$

Since we defined ξ such that $\xi^b_{i,i} = 1$, we can write:

$$L^b((c^b, a^b), (c^{-b}, a^{-b})) = \textstyle\sum_{i \in \mathcal{S}} \gamma^b_i \sigma^2_i(\mu_i(\boldsymbol{a}_i)) + \eta^b \sum_{i \in \mathcal{S}} \left(c^b_i - a^b_i \left[\sum_{l \in \mathcal{S}} \xi^b_{i,l} \sigma^2_l(\mu_l(\boldsymbol{a}_l)) \right] \right)$$

Similarly, the expected payment for any data source s and data aggregator b is given by:

$$\mathrm{p}^b_s((c^b_s, a^b_s), e) = c^b_s - a^b_s \left(\textstyle\sum_{i \in \mathcal{S}} \xi^b_{s,i} \sigma^2_i(e_i) \right)$$

Before proceeding, we provide an interpretation of the constants introduced above. The constant β^b_s denotes the relevance of data sampled from the point x_s when constructing aggregator b's estimate, given the distribution of all of the data sources. The parameter γ^b_s corresponds to the level of demand that aggregator $b \in \mathcal{B}$ has for high-quality data from source $s \in \mathcal{S}$, factoring in the benefit this data supplies to the competitors of b. In other words, γ parameters capture the effects of the non-rivalrous nature of data. The parameter $\xi^b_{s,l}$ denotes a measure of coupling that exists between the payment contracts p^b_s and p^b_l. In the case of a single aggregator (i.e. [17]), this coupling did not prove problematic. In contrast, when there are multiple aggregators, each aggregator has an incentive to try and exploit this coupling, as shall become clear in our ensuing analysis. This coupling will play a central role in determining the existence and efficiency of equilibrium behavior in the data market.

Collecting the various expressions we have introduced, aggregator b's optimization problem can be re-written as:

$$\min_{(c^b, a^b)} \; L^b((c^b, a^b), (c^{-b}, a^{-b})) \tag{20}$$

$$\text{s.t.} \sum_{j \in \mathcal{B}} \left[c^j_s - a^j_s \left(\sum_{i \in \mathcal{S}} \xi^j_{s,i} \sigma^2_i(\mu_i(\boldsymbol{a}_i)) \right) \right] \geq \mu_s(\boldsymbol{a}_s), \;\; \forall s \in \mathcal{S}$$

$$c^b_s - a^b_s \left(\sum_{i \in \mathcal{S}} \xi^b_{s,i} \sigma^2_i(\mu_i(\boldsymbol{a}_i)) \right) \geq 0, \;\; \forall s \in \mathcal{S}$$

$$\boldsymbol{a}_s \in \mathcal{A}_s, \;\; \forall s \in \mathcal{S} \qquad a^b_s \geq 0, \;\; \forall s \in \mathcal{S}$$

Without loss of generality, we let $\eta^b = 1$, by normalizing the γ^b_s accordingly. Note that the constraint $c^b_s \geq 0$ can be omitted, in light of the constraint $c^b_s - a^b_s \left(\sum_{i \in \mathcal{S}} \xi^b_{s,i} \sigma^2_i(e^*_i) \right) \geq 0$, since each $\xi^b_{s,i} \geq 0$ and $a^b_s \geq 0$.

IV. Generalized Nash Equilibria in the Data Market

It is important to note that the constraints each aggregator faces in her optimization problem (20) depend on the actions taken by the rest of the aggregators in the data market. In particular, in order to ensure that the IR and IC constraints are maintained in equilibrium, we require an equilibrium concept which allows each aggregator's admissible action space to depend on the choice of contract parameters selected by the other aggregators in the data market. Thus, we will employ the notion of a *generalized Nash equilibrium* [23] to study competitive outcomes in the data market, which is a natural extension of the typical notion of Nash equilibrium to this setting.

Let $\mathcal{Z}_b \subset \mathbb{R}^{2N}$ be aggregator b's actions space; that is, $z^b = (c^b, a^b) \in \mathbb{R}^{2N}$ where $c^b = (c^b_s)_{s \in \mathcal{S}}$ and $a^b = (a^b_s)_{s \in \mathcal{S}}$. Each aggregator $b \in \mathcal{B}$ solves a parametric nonlinear programming problem given by

$$P_b(z^{-b}) := \min_{z^b} \; \{ L^b(z^b, z^{-b}) : z^b \in \mathcal{M}^b(z^{-b}) \} \tag{21}$$

where $\mathcal{M}^b(z^{-b}) = \{ z^b : \; k^b_j(z^b, z^{-b}) \geq 0, \;\; \forall \; j \in \mathcal{J}^b \} \subset \mathcal{Z}_b$ with $\mathcal{J}^b = \{1, \ldots, |\mathcal{J}^b|\}$ a finite set indexing the constraint functions of aggregator b. Note that, unlike in the classic definition of a Nash equilibrium, the admissible action space of aggregator b depends on z^{-b}, the actions of $-b = \mathcal{B} \setminus \{b\}$.

We say $\{P_b(\cdot)\}_{b \in \mathcal{B}}$ is a generalized Nash (GN) equilibrium problem. A GN equilibrium is defined as follows.

Definition 4. *A point $z = (z^1, \ldots, z^{|\mathcal{B}|}) \in \prod_{b=1}^{|\mathcal{B}|} \mathcal{Z}_b$ is said to be a GN equilibrium for $\{P_b(\cdot)\}_{b \in \mathcal{B}}$ if for all $b \in \mathcal{B}$, z^b solves $P_b(z^{-b})$.*

We now analyze the game between the aggregators utilizing the notion of a GN problem and GN equilibrium. We will characterize the existence and uniqueness of GN equilibria in two scenarios. In Section IV-A, we will consider the case where the effort spaces of data sources are unbounded, i.e. $\mathcal{E}_s = \mathbb{R}_{\geq 0}$. In Section IV-B, we will characterize the case where each data source has an upper bound on the level of effort they can exert, i.e. $\mathcal{E}_s = [0, e^{max}_s)$. In Section IV-C, we will then address the social efficiency of the equilibria identified in Section IV-A. A similar analysis of the equilibria identified in Section IV-B can be found in Appendix B.

Before preceding to out main results, we provide a technical lemma that will have a central role in our ensuing analysis and introduce some notation which will simplify the statement of our results. For compactness, for a given set of a parameters we define $\mu(a) = (\mu_s(\boldsymbol{a}))_{s \in \mathcal{S}}$. (Recall that $\boldsymbol{a}$ is the sum of a parameters, as defined in Equation (13).)

Lemma 2. *Suppose $z = (z^b)_{b \in \mathcal{B}}$, where $z^b = (c^b, a^b)$, is a GN equilibrium for the game $\{P_b(\cdot)\}_{b \in \mathcal{B}}$ defined by* (20)*. Then for each $s \in \mathcal{S}$:*

$$\textstyle\sum_{j \in \mathcal{B}} c^j_s - \sum_{j \in \mathcal{B}} a^j_s \big(\sum_{i \in \mathcal{S}} \xi^j_{s,i} \sigma^2_i(\mu_i(\boldsymbol{a}_i)) \big) - \mu_s(\boldsymbol{a}_s) = 0 \tag{22}$$

In other words, the IR constraint is always binding in equilibrium, and the expected payment to data source s is equal to the effort exerted in equilibrium: $\mathrm{p}_s((c_s, a_s), \mu(a)) = \mu_s(\boldsymbol{a}_s)$

Proof. Suppose that there is an equilibrium in which the IR constraint is not binding for some data source s. Then, there must a exists an aggregator b whose non-negativity constriant corresponding to source s is also not binding. Thus, this cannot be an equilibrium as aggregator b can unilaterally improve their payoff by decreasing c^b_s without causing any of the constraints to be violated, contradicting the assertion that the given selection of parameters is an equilibrium. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

Notation	Meaning	Defined or First Used in Equation
s	index of data source	–
$\mathcal{S}$	index set of data sources	–
b	index of aggregator	–
$\mathcal{B}$	index set of aggregators	–
p_s^b	expected payment from aggregator b to source s	(7)
a_s^b	linear term in p_s^b; used to adjust level of effort e_s in equilibrium	(6)
a_s	vector containing the a parameters offered to source s by the members of $\mathcal{B}$	–
c_s^b	constant term in p_s^b; used to ensure incentive compatibility in equilibrium	(6)
c_s	vector containing the c parameters offered to sources s by the members of $\mathcal{B}$	–
$\boldsymbol{a}_s$	sum of a parameters offered to source s across all members of $\mathcal{B}$	(13)
$\underline{\boldsymbol{a}}_s$	mimimum value of $\boldsymbol{a}_s$ required to ensure source s does not exert negative effort	(15)
$\overline{\boldsymbol{a}}_s$	minimum value of $\boldsymbol{a}_s$ at which data source s exerts her maximum effort	(14)
$\mathcal{A}_s$	$[\underline{\boldsymbol{a}}_s, \overline{\boldsymbol{a}}_s]$, the allowable range of $\boldsymbol{a}_s$	–
μ_s	implicit map which returns the equilibrium value of e_s as a function of $\boldsymbol{a}_s$	–
ζ_j^b	level of competition between $j, b \in \mathcal{B}$	(5)
β_s^b	relevance of data from x_s in constructing aggregator b's estimator	(17)
γ_s^b	aggregate demand for e_s from b	(19)
$\boldsymbol{\gamma}_s$	sum of demand for data source s across all members of $\mathcal{B}$	(23)
$\xi_{s,l}^b$	coupling between p_s^b and p_l^b	(36)

TABLE I: Notation Reference Chart

The result of Lemma 2 is a well-known result in contract design—that is, the individual rationality constraint always binds for the optimal contract [24]. As shall become clear in our analysis in the following sections, the equality (22) forms an implicit constraint that appears in each of the aggregators' optimizations, which will be directly responsible for the degeneracy observed in the data market. Roughly speaking, while the a parameters selected by the aggregators determine the level of effort that the data sources will exert, the c parameters determine what portion of this effort each aggregator is expected to compensate.

For each $s \in \mathcal{S}$, define:

$$\boldsymbol{\gamma}_s = \textstyle\sum_{j \in \mathcal{B}} \gamma_s^j \tag{23}$$

which can be interpreted to be the total demand for high quality data from data source s. Next, we define:

$$\boldsymbol{c}_s = \sum_{j \in \mathcal{B}} c_s^j \qquad q_s^b(a) = a_s^b \left(\sum_{i \in \mathcal{S}} \xi_{s,i}^j \sigma_i^2(\mu_i(\boldsymbol{a}_i)) \right)$$

$$q_s(a) = \sum_{j \in \mathcal{B}} q_s^j(a) + \mu_s(\boldsymbol{a}_s)$$

Note that Lemma 2 implies that if (c, a) is an GN equilibrium in the game between the aggregators then $\boldsymbol{c}_s = q_s(a)$ will hold for each $s \in \mathcal{S}$. Moreover, the non-negativity constraints in the game between the buyers will hold only if $c_s^b \geq q_s^b(a)$ for each $s \in \mathcal{S}$ and $b \in \mathcal{B}$.

A. Unbounded Effort Spaces

Let us first consider the case where there is no upper bound on the effort the data sources may exert, i.e. $\mathcal{E}_s = \mathbb{R}_{\geq 0}$.

Theorem 1. *Consider the game $\{P_b(\cdot)\}_{b \in \mathcal{B}}$, where each aggregator's objective is to solve the optimization in* (20)*. Suppose that for each $s \in \mathcal{S}$, $\mathcal{E}_s = \mathbb{R}_{\geq 0}$ and $\boldsymbol{\gamma}_s \geq \underline{\boldsymbol{a}}_s$. Further, suppose that $\gamma_i^j > 0$, $\forall\ i \in \mathcal{S}, j \in \mathcal{B}$. Then, there is either no GN equilibrium or an infinite number of GN equilibria. Moreover, if $(\bar{c}, \bar{a})$ is a GN equilibrium, then the following conditions hold:*

1) The set of infinite GN equilibria is given by:

$$\left\{(c, a) : a = \bar{a},\ \boldsymbol{c}_s = q_s(\bar{a}),\ c_s^b \geq q_s^b(\bar{a}),\ \forall s, \forall b\right\}$$

That is, the a parameters selected by the aggregators are the same across each GN equilibrium, and all degeneracy lies in the equilibrium c parameters which lie in the $|B|$-dimensional convex polytope defined above.

2) The effort exerted by each data source is the same in each GN equilibrium and the efforts constitute a unique induced dominated strategy equilibrium between the data sources. More precisely, each data source exerts effort $\mu_s(\bar{\boldsymbol{a}}_s)$ in all GN equilibria.

Before going ahead with the proof of the theorem, we discuss its hypotheses and implications. The hypothesis that $\boldsymbol{\gamma}_s \geq \underline{\boldsymbol{a}}_s$ implies that there is enough demand for the data from source s such that she does not have incentive to exert negative effort in equilibrium. Together, the aggregators will provide sufficient incentive to s so that s accepts each of the contracts offered to her, and truthfully report her query-response. When we investigate the case where s only provides readings to a subset of the aggregators in Section V, only the relevant subset of aggregators must maintain this constraint. This condition places a restriction on what subsets of incentives from the aggregators each data source is willing to accept.

As we discovered in Section III-F, the $a = (a^b)_{b \in \mathcal{B}}$ parameters selected by the aggregators uniquely determine how much effort the data sources exert in equilibrium. Intuitively, the fact that the a parameters are constant across all GN equilibria means that, when GN equilibria do exist in the game between the aggregators, the aggregators have agreed to incentivize the data sources to each exert a particular level of effort. The proof of the theorem will shed some light on how this unique choice of a parameters is selected when GN equilibria exist, and also demonstrate what 'goes wrong' in cases where the aggregators cannot agree on how much effort to incentivize the sources to exert. In the latter case, no GN equilibrium solution exists in the game between the aggregators. Further commentary on this point is provided after the proof of Theorem 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

Meanwhile, for a fixed profile of a parameters, the $c = (c^b)_{b\in\mathcal{B}}$ parameters determine how much of this effort each aggregator is responsible for compensating in expectation. Even when aggregators are able to agree on how much effort to incentivize from the data sources and select the unique GN equilibria choice for a, there is a non-uniqueness in the c parameters in equilibrium. This implies that there is a fundamental ambiguity in who will fund the exertion of the data sources. In the extreme case, it is possible for one aggregator to pay for the entirety of the expected compensation offered to the data sources, while the other aggregators pay nothing in expectation.

Proof of Theorem 1. By Lemma 2, we have that:

$$c_s^b = \sum_{j\in\mathcal{B}} a_s^j \left(\sum_{i\in\mathcal{S}} \xi_{s,i}^j \sigma_i^2(\mu_i(\boldsymbol{a}_i))\right) + \mu_s(\boldsymbol{a}_s) - \sum_{j\in -b} c_s^j \tag{24}$$

Plugging in this constraint, the cost function for aggregator b can be expressed as:

$$\tilde{L}^b(a^b,(c^{-b},a^{-b})) = \sum_{i\in\mathcal{S}} \Big(\gamma_i^b \sigma_i^2(\mu_i(\boldsymbol{a}_i)) + \sum_{j\in -b}\Big[a_i^j\left(\sum_{l\in\mathcal{S}} \xi_{i,l}^j \sigma_l^2(\mu_l(\boldsymbol{a}_l))\right) - c_i^j\Big] + \mu_i(\boldsymbol{a}_i)\Big)$$

By swapping the roles of i and l in the middle term above, aggregator b's cost can be decomposed into the sum of costs for each data sources. We define:

$$\tilde{L}_s^b(a_s^b,(c^{-b},a^{-b})) = \big(\gamma_s^b + \sum_{j\in -b}\sum_{l\in\mathcal{S}} a_l^j \xi_{l,s}^j\big)\sigma_s^2(\mu_s(\boldsymbol{a}_s)) - \sum_{j\in -b} c_s^j + \mu_s(\boldsymbol{a}_s)$$

Then aggregator b's optimization problem reduces to:

$$\begin{aligned}\min_{a^b}\ & \sum_{s\in\mathcal{S}} \tilde{L}_s^b(a_s^b,(c^{-b},a^{-b}))\\ \text{s.t.}\ & \sum_{j\in -b}\left[a_s^j\left(\sum_{i\in\mathcal{S}} \xi_{s,i}^j \sigma_i^2(\mu_i(\boldsymbol{a}_i))\right) - c_s^j\right] + \mu_s(\boldsymbol{a}_s) \geq 0\\ & \boldsymbol{a}_s \in \mathcal{A}_s,\ a_s^b \geq 0\end{aligned}$$

Note that the cost does not depend on c_s^b, for any s. We complete the argument by ignoring the constraints and showing that the constraints are satisfied for the set of equilibria we characterize.

Differentiating the cost with respect to a_s^b and applying (16) and $\xi_{s,s}^j = 1$ for all j, we have that:

$$D_{a_s^b}\tilde{L}_s^b = \tfrac{1}{\boldsymbol{a}_s}\left(a_s^b - \gamma_s^b - \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j\right) D_{\boldsymbol{a}_s}\mu_s$$

where $D_x \equiv \frac{\partial}{\partial x}$. Applying Lemma 1, which states $D_{\boldsymbol{a_s}}\mu_s > 0$, we get the following conditions:

$$\begin{cases} D_{a_s^b}\tilde{L}_s^b < 0, & \text{if } 0 \leq a_s^b < \gamma_s^b + \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j \\ D_{a_s^b}\tilde{L}_s^b = 0, & \text{if } a_s^b = \gamma_s^b + \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j \\ D_{a_s^b}\tilde{L}_s^b > 0, & \text{if } a_s^b > \gamma_s^b + \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j \end{cases}$$

Hence, the a_s^b that minimizes aggregator b's cost satisfies:

$$a_s^b = \gamma_s^b + \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j \tag{25}$$

Performing this analysis for all combinations of $s \in \mathcal{S}$ and $b \in \mathcal{B}$ yields a system of $M \times N$ equations with $M \times N$ unknowns, of the form (25).

Let a denote a column vector with entries a_i^j and let γ denote a column vector containing all the terms of the form γ_i^j for each $i \in \mathcal{S}$ and $j \in \mathcal{B}$. Then, (25) can be written as the system of equations given by:

$$a = \Xi a + \gamma \tag{26}$$

Here, Ξ is a non-negative matrix whose entries are composed of the $\xi_{i,l}^j$ values such that (26) expresses the set of equality constraints defined by (25) for all $s \in \mathcal{S}$ and $b \in \mathcal{B}$. To solve this reduced game, it suffices to find a solution to (26) such that $\boldsymbol{a}_i \in \mathcal{A}_i$ and $a_i^j \geq 0$ for all $i \in \mathcal{S}$ and $j \in \mathcal{B}$.

Let us consider first solutions to the system of equations (26). Systems of equations of this form are well studied in the economics literature, as they are of the form specified by the celebrated Leontief input-output model. It has been shown that such systems of equations have a non-negative solution if and only if $\rho(\Xi) < 1$, where $\rho(\Xi)$ is the spectral radius of Ξ [25]. Moreover, if such a solution exists, it must be unique.

Thus, if $\rho(\Xi) < 1$, inversion of the $(I - \Xi)$ matrix yields the equilibrium a, and, by Lemma 2, we can pick any c such that for each $s \in \mathcal{S}$, $b \in \mathcal{B}$,

$$c_s^b \geq a_s^b\left(\sum_{l\in\mathcal{S}} \xi_{s,l}^b \sigma_l^2(\mu_l(\boldsymbol{a}_l))\right) \tag{27}$$

and (24) hold.

If $\rho(\Xi) \geq 1$, there will not exist a non-negative solution and there is no point (c, a) that simultaneously optimizes (20) for all $b \in \mathcal{B}$. Finally, we demonstrate that none of the solutions to (26) and corresponding c values defined above violate the constraint $\boldsymbol{a}_s \in \mathcal{A}_s = [\underline{\boldsymbol{a}}_s, \infty)$. By inspecting equations of the form (25), we see that $\boldsymbol{a}_s \geq \boldsymbol{\gamma}_s \geq \underline{\boldsymbol{a}}_s$, and thus the constraints remain satisfied. It follows that there is either a unique set of a parameters defining a GN equilibria for the reduced game between the aggregators, otherwise there is no GN equilibrium. Moreover, in the case that an equilibrium choice of a does exist, by inspection we see that the polytope of c parameters defined in the statement of the theorem constitute GN equilibria for the full game. □

It is interesting to note that the existence of GN equilibria depends solely on the value of the $\xi_{s,l}^b$ parameters; it does not depend on the magnitude of the γ_s^b parameters (given that they are large enough to ensure participation of all parties). This implies that the existence of GN equilibria follows from the form of the contract mechanisms, and does not depend on whether or not there are solutions that are beneficial to all parties involved.

B. Bounded Effort Spaces

Let us now consider the case where the data sources' effort space is upper-bounded, i.e. $\mathcal{E}_s = [0, e_s^{\max}]$, $0 \leq e_s^{\max} < \infty$.

Theorem 2. *Consider the game $\{P_b(\cdot)\}_{b\in\mathcal{B}}$ where each aggregator's objective is to solve the optimization in* (20)*. Suppose that for each $s \in \mathcal{S}$, $\mathcal{E}_s = [0, e_s^{\max}]$ with $0 \leq e_s^{\max} < \infty$ and that $\gamma_i^j > 0$ for all $i \in \mathcal{S}, j \in \mathcal{B}$. There is an infinite number of GN equilibria $z = (z^b)_{b\in\mathcal{B}}$. Moreover, the following statements hold:*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

1) *There may exist two GN equilibria* (c_1, a_1) *and* (c_2, a_2) *such that* $a_1 \neq a_2$*;*
2) *If* $(\bar{c}, \bar{a})$ *constitutes a GN equilibrium then the following set of parameters also constitute GN equilibria:*

$$\big\{(c,a) : a = \bar{a},\ \forall s \in \mathcal{S}, b \in \mathcal{B},\ c_s = q_s(\bar{a}),\ c_s^b \geq q_s^b(\bar{a})\big\}$$

3) *While the data sources may exert different levels of effort across different equilibria, each collection of* a *parameters chosen by the aggregators still induce a dominanted strategy equilibrium between the data sources.*

Proof. Following the proof of Theorem 1, the problems $P_b(\cdot)$ can be reduced to the optimization problem

$$\tilde{P}_b(z^{-b}) := \min_{a^b}\ \big\{\tilde{L}^b(a^b, z^{-b}) : \forall s \in \mathcal{S}, \boldsymbol{a}_s \in \mathcal{A}_s,\ a^b \geq 0\big\}$$

To show existence, we show that the game defined by $\{\tilde{P}_b(\cdot)\}_{b\in\mathcal{B}}$ satisfies the assumptions of Theorem 5, which is originally from [26] and can be found in the Appendix.

First, we note that the objective function of each buyer is continuous in each of its arguments. Indeed, Assumption 3 ensures that σ_s^2 is continuous and Lemma 1 ensures the map μ_s is continuous for each $s \in \mathcal{S}$. Continuity of the objective function follows by recalling that the composition of continuous maps yields a continuos map. Next, for each $j \in \mathcal{B}$, the constraints defining $\mathcal{M}_{-j}(z^{-j})$ are continuous, and thus the correspondence $z^{-j} \mapsto \mathcal{M}_{-j}(z^{-j})$ is upper semi-continuous (indeed, even continuous). Continuing with the analysis from Theorem 1, it is easy to see that in the case where we constrain $\boldsymbol{a}_i \in \mathcal{A}_i$, the best response set for aggregator b is defined by the following conditional statements for each $s \in \mathcal{S}$:

$$\begin{cases} a_s^b = \underline{\boldsymbol{a}}_s - \sum_{j\in -b} a_s^b & \text{if } T_s^b(a^{-b}) < \underline{\boldsymbol{a}}_s \\ a_s^b = \gamma_s^b + \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j & \text{if } T_s^b(a^{-b}) \in \mathcal{A}_s \\ a_s^b = \overline{\boldsymbol{a}}_s - \sum_{j\in -b} a_s^j & \text{if } T_s^b(a^{-b}) > \overline{\boldsymbol{a}}_s \end{cases} \quad (28)$$

where $T_s^b(a^{-b}) = \gamma_s^b + \sum_{j\in -b}\sum_{l\in -s} a_l^j \xi_{l,s}^j + \sum_{j\in -b} a_s^j$.

Thus, the best response set for aggregator b is always a singleton which in turn implies it is always contractable. Further, each best response mapping $\mathrm{BR}_b\colon a^{-b} \mapsto a^b$ is continuous in a^{-b}. Hence,

$$\min_{a^b} L^b(a^b, a^{-b}) = L^b(\mathrm{BR}_b(a^{-b}), a^{-b}) \quad (29)$$

and thus the mapping $a^{-b} \mapsto \min_{a^b} L^b(a^b, a^{-b})$ is continuous since L^b and BR_b are continuous and the composition of continuous maps is continuous. Thus, by Theorem 5 there exists a GN equilibrium for the reduced game between the aggregators. As in the proof of Theorem 1, we see that for any collection of a parameters that constitute a GN equilibrium for this simplified game, any collection of c parameters that lie in the convex polytope defined in the statement of the theorem constitute a GN equilibrium for the full game. □

We now remark on why there are always GN equilibria to the game between the aggregators when each source can exert a finite amount of effort (i.e. Theorem 2), but there may not be a GN equilibrium when the sources are allowed to exert infinite effort (i.e. Theorem 1). Referring to (26), consider the case where $\rho(\Xi) = k < 1$ and GN equilibria exist. Suppose we replace Ξ in (26) with $\alpha\Xi$, where $\alpha > 1$, and note that $\rho(\alpha\Xi) = \alpha\rho(\Xi)$. As α is increased towards $\frac{1}{k}$, the matrix $(I - \alpha\Xi)$ gets closer to becoming singular, and the corresponding solution to the system of equations, a, approaches infinity. This implies the data sources exert an infinite amount of effort in equilibrium.

Intuitively, this corresponds to the coupling between the payment contracts approaching some critical limit, past which point no GN equilibria exist. However, in the case where the data sources are constrained to exert a finite amount of effort, this run-away behavior is not possible, and GN equilibria always exist. While the constraints bounding the effort the data sources may exert ensure the existence of GN equilibria, their activation and inactivation may lead to a degeneracy in how much effort each data source exerts in equilibrium.

Note that, in Theorem 1 and Theorem 2, we assume that either all of the data sources are constrained in their effort, or all unconstrained in their effort. In the case where some data sources are constrained and others unconstrained, the task of determining existence of equilibria to the game becomes a combinatorial endeavor. We exclude the analysis of this case, as it is largely an algebraic exercise and lends little insight to the broader problem.

C. Conditions for Social Inefficiency

The question of equilibrium *efficiency* or *quality* arises naturally in game theoretic settings. In this section, we identify necessary and sufficient conditions under which the equilibria are socially inefficient; as we will see in Theorem 3, as soon as any non-diagonal ξ parameter is non-zero, there will be social inefficiency.

Due to the wide variety of estimators data aggregators can use, as well as effort-to-variance functions that characterize data sources, it is difficult to provide interesting general bounds on the price of anarchy, a widely used metric for the inefficiency of equilibria [27]. However, when both are specified, the price of anarchy can be explicitly calculated [22].

In this section, we will focus our attention on the case where the data sources are unconstrained in the effort they exert, i.e. when $\mathcal{E}_s$ is unbounded, as the results in this case provide a clearer intuition for how our chosen class of mechanisms give rise to inefficiencies. However, in the interest of completeness, in Appendix B the result is extended to the case where data sources are effort-constrained.

Let us denote by $\boldsymbol{e}$ the vector denoting the level of effort the data sources exert. The social cost is defined as the sum of the cost experienced by all parties.

Definition 5 (Ex-ante Social Cost)**.** *Suppose that* $\eta^j = 1$ *for each aggregator* $j \in \mathcal{B}$*. We define the ex-ante social cost to be the sum of the utility functions of all the data aggregators and data sources—that is:*

$$\mathcal{L}(\boldsymbol{e}) = \sum_{j\in\mathcal{B}} \big(\mathbb{E}\big[\big(\hat{f}^j_{\mathcal{X}_j}(x^*) - f(x^*)\big)^2 - \sum_{k\in -j} \zeta_k^j \big(\hat{f}^k_{\mathcal{X}_k}(x^*) - f(x^*)\big)^2\big]\big) + \sum_{s\in\mathcal{S}} e_s \quad (30)$$

Note that this sum does not include any of the payments made in the marketplace, as they are simply lossless transfers of wealth. The fact that $\eta^j = 1$ for each $j \in \mathcal{B}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

ensures that these transfers of wealth are lossless from a utility perspective—i.e. the aggregators and sources value the payment equally. However, this is simply a rescaling and, more importantly, it allows us to isolate the social loss due to the mechanism, and ignore any losses due to differential preferences in payment currency. Indeed, note that the social cost only depends on the effort exerted by the sources.

The following result states that there is always a unique level of effort that minimizes the social cost.

Lemma 3. *Suppose that $\gamma_i^j > 0$, $\forall j \in \mathcal{B}$ and $\forall i \in \mathcal{S}$. There is a unique minimizer of $\mathcal{L}(\boldsymbol{e})$.*

Proof. The ex-ante social cost can be re-written as $\mathcal{L}(\boldsymbol{e}) = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{B}}\gamma_i^j\sigma_i^2(e_i) + \sum_{i\in\mathcal{S}} e_i$. By Assumption 3 and the assumption that $\gamma_s^j > 0$ for some $j \in \mathcal{B}$, $D_{e_s}^2\mathcal{L}(\boldsymbol{e})$ is strictly positive. In addition, $D_{e_s}\mathcal{L}(\boldsymbol{e})$ does not depend on e_i for $i \in \mathcal{S}\setminus\{s\}$. Hence, $D_{\boldsymbol{e}}^2\mathcal{L}$ has positive entries on the diagonal and entries of zero everywhere else so that it is positive definite which, in turn, implies $\mathcal{L}$ is strictly convex. Thus, since $\mathcal{E}$ is a convex set, $\mathcal{L}$ has a unique minimizer on $\mathcal{E}$. □

The price of anarchy is defined for each Nash equilibrium as the ratio of the social cost under the Nash equilibrium to the socially optimal cost.

Definition 6 (Ex-ante Price of Anarchy)**.** *The ex-ante price of anarchy is given by* $\mathrm{PoA}(\boldsymbol{e}) = \frac{\mathcal{L}(\boldsymbol{e})}{\mathcal{L}(\hat{\boldsymbol{e}})}$*, where $\hat{\boldsymbol{e}} \in \mathcal{E}$ is the minimizer of $\mathcal{L}$.*

Since $\hat{\boldsymbol{e}}$ is unique minimizer of $\mathcal{L}$, for all $\boldsymbol{e} \neq \hat{\boldsymbol{e}}$ we must have that $\mathrm{PoA}(\boldsymbol{e}) > 1$. Intuitively, the larger $\mathrm{PoA}(\boldsymbol{e})$, the more socially inefficient the solution is. With this metric in mind, we provide necessary and sufficient conditions for the game between the aggregators to yield a socially efficient outcome, in the case that the effort space for each data source is unbounded.

Theorem 3. *Suppose $\gamma_i^j > 0$, $\forall i \in \mathcal{S}$, $\forall j \in \mathcal{B}$. Further suppose that $\mathcal{E}_s = [0,\infty)$ for each $s \in \mathcal{S}$. Then, there exists a GN equilibrium to the game $\{P^b(\cdot)\}_{b\in\mathcal{B}}$ between the aggregators for which the price of anarchy is equal to one if and only if for each $j \in \mathcal{B}$ and each $i,l \in \mathcal{S}$ such that $i \neq l$ we have that $\xi_{i,l}^j = 0$.*

Proof. When there is no GN equilibria of the aggregators' game, the proof is trivial. On the other hand, when there is GN equilibria we have that

$$D_{e_s}\mathcal{L}(\boldsymbol{e}) = 2\boldsymbol{\gamma}_s\sigma_s(\hat{e}_s)\tfrac{d}{de_s}\sigma_s(\hat{e}_s) + 1 = 0. \tag{31}$$

Since $\mathcal{L}$ is strictly convex, solving the first order conditions in (31) yields the global minimizer. Just as with (16), the solution to (31), $\hat{e}_s \in \mathbb{R}_+$, is implicitly defined by $\hat{\mu}_s : \boldsymbol{\gamma}_s \mapsto \hat{e}_s$.

Moreover, at a GN equilibrium, we have $2\boldsymbol{a}_s\sigma_s(e_s)\frac{d}{de_s}\sigma_s(e_s) + 1 = 0$. Thus, as a consequence of Lemma 1, the data sources exert $\hat{e}_s \in \mathbb{R}_+$ if and only if $\boldsymbol{a}_i = \boldsymbol{\gamma}_i$ for all $i \in \mathcal{S}$. By the proof of Theorem 1 an equilibrium choice of the parameters (c,a) must satisfy:

$$\boldsymbol{a}_i = \boldsymbol{\gamma}_i + \textstyle\sum_{j\in\mathcal{B}}\sum_{k\in-j}\sum_{l\in-i} a_l^k\xi_{l,i}^k \tag{32}$$

Thus, we will have $\boldsymbol{a}_i = \boldsymbol{\gamma}_i$ if and only if:

$$0 = \textstyle\sum_{j\in\mathcal{B}}\sum_{k\in-j}\sum_{l\in-i} a_l^k\xi_{l,i}^k \tag{33}$$

However, $a_s^b > \gamma_s^b > 0$ for some $s \in \mathcal{S}$ and $b \in \mathcal{B}$ since in equilibrium we have $a_b^s = \gamma_s^b + \sum_{j\in-b}\sum_{l\in-s} a_l^j\xi_{l,s}^j$ (see (28)), and by assumption $\xi_{l,i}^j > 0$ for some $j \in \mathcal{B}$ and some $i,l \in \mathcal{S}$ with $i \neq l$. Thus, (33) cannot hold. □

Theorem 3 confirms the typical result that Nash equilibria are generally (ex-ante) inefficient. However, it further shows that, in the particular case of this framework, GN equilibria are efficient only when there is no coupling between the payments the aggregators make to data sources. In this light, we find it appealing to regard the data sources as public good which the aggregators have incentive to exploit. As the following result demonstrates, this problematic coupling will always arise when the aggregators utilize linear regression.

Corollary 1. *Suppose that each aggregator $b \in \mathcal{B}$ has estimator $\hat{f}^b$ which is linear regression. Further suppose that the conditions of Theorem 3 hold. Then there does not exist a GNE solution to $\{P^b(\cdot)\}_{b\in\mathcal{B}}$ wherein the price of anarchy is equal to* 1.

Proof. It can be shown (see [17, Footnote 6]) that

$$g_b(x_{-s},\delta_{x_s},\sigma_{-s}^2(e_{-s})) = \mathbb{E}_{\tilde{x}\sim\delta_{x_s}}\big[[\tilde{x}^T,1]\cdot(X^TX)^{-1}X^T \cdot \mathrm{diag}(\sigma_{-s}^2(e_{-s}))\cdot X(X^TX)^{-1}\cdot[\tilde{x}^T,1]^T\big]$$

where X is the matrix whose rows are $[x_i^T,1]$ for $i \in -s$ and $\mathrm{diag}(\sigma_{-s}^2(e_{-s}))$ is the diagonal matrix whose (i,i)–th entry is $\sigma_i^2(e_i)$. By inspection, $g_b(x_{-s},\delta_{x_s},\sigma_{-s}^2(e_{-s}))$ is ill-defined if $[x_s^T,1]$ is orthogonal to $\mathrm{span}\{([x_i^T,1])_{i\in-s}\}$. This, in turn, implies the payment contract p_s^b is ill-defined. Since all the payments are assumed well-defined, $[x_s^T,1]$ cannot be orthogonal to $\mathrm{span}\{([x_i^T,1])_{i\in-s}\}$. Thus by inspection there exists $i,l \in \mathcal{S}$ such that $i \neq l$ and $\xi_{i,l}^b > 0$. Thus, by Theorem 3, there is no efficient equilibrium. □

V. Partial Data Sharing

Thus far we have made the restrictive assumption that each of the data sources accepts payment from and provides query responses to each of the aggregators in the data market. We have done this primarily to ease the introduction of the notation needed to state and prove our main results. In this section, we remove this assumption and analyze the case where each of the data sources only accepts payment contracts from and provides query responses to a subset of the data aggregators. That is, we now assume that prior to the first stage outlined at the beginning of Section III each data source $s \in \mathcal{S}$ has agreed to accept the incentives issued by some subset of the buyers $\mathcal{B}_s \subset \mathcal{B}$. As we shall see, these changes do not alter our previous analysis significantly. When aggregator $b \in \mathcal{B}$ only purchases data from a subset of the data sources, the primary difference is that b now has fewer contract parameters (c^b,a^b) and fewer IR and non-negativity constraints in his optimization. The removal of these terms and constraints reduces the dimensionality of the degeneracy observed in the equilibria of the datamarket, but the overall structure of our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

analysis changes little. In order to demonstrate this point, in this section we will focus on demonstrating in some detail how a result analogous to Theorem 1 can be obtained in this setting.

In practice, the mechanism by which data sources choose which incentives to accept and which aggregators to work with could be quite complicated. Although this is an important and interesting point for future research, in this paper we will assume that the sets $\mathcal{B}_s$ are exogenously given for all s. Our purpose here is to show that for any exogenously fixed assignments of data sources to data aggregators, the degeneracies we highlighted earlier remain whenever at least one data source receives payment from and provides data to multiple aggregators, i.e. whenever the non-rivalrous nature of data has an effect on the data market.[9] Throughout the section we will outline how the formulation we have considered must be modified to fit this more general setting, and discuss how the results we have presented thus-far carry through. As we shall see, the generalization is rather straightforward, and thus some details are omitted in the interest of brevity.

Once it has been decided which data sources will provide data to which aggregators, the interactions of the data market proceed as before. Each aggregator $b \in \mathcal{B}$ issues incentives $p^b = (p_s^b)_{s\in\mathcal{S}_b}$ of the form (6) to the members of $\mathcal{S}_b$, then each data source $s \in \mathcal{S}$ evaluates the payments $p_s = (p_s^j)_{j\in\mathcal{B}_s}$, decides what level of effort to exert when producing y_s and then shares this reading with the members of $\mathcal{B}_s$. Each aggregator then processes the data she has received to construct her estimate for f_b, issues payments p^b, and incurs loss L^b.

Note that we have abused notation in redefining p_s and p^b above to only reflect the subset of payments that are issued in this section. Similar abuses will follow as we redefine a number of objects from earlier in the document to be appropriate for this new setting. Roughly speaking, each of these items will be redefined by replacing $\mathcal{B}$ with $\mathcal{B}_s$ and $\mathcal{S}$ with $\mathcal{S}_b$ where appropriate. For example, the buyers now need only to select the parameters $c^b = (c_s^b)_{s\in\mathcal{S}_b}$ and $a^b = (a_s^b)_{s\in\mathcal{S}_b}$ when issuing incentives. We will omit the details of some of these changes when context makes our meaning clear.

We may now model the utility for each data source s by:

$$u_s(e_s, p_s) = \mathbb{E}\Big(\sum_{j\in\mathcal{B}_s} p_s^j(y^j(e))\Big) - e_s \tag{34}$$

and the loss for each aggregator b by:

$$\begin{aligned} L^b(p^b, e) = \mathbb{E}\Big[&\big(\hat{f}^b_{\mathcal{X}_b}(x^*) - f(x^*)\big)^2 \\ &- \textstyle\sum_{j\in-b} \zeta_j^b \big(\hat{f}^j_{\mathcal{X}_j}(x^*) - f(x^*)\big)^2 \\ &+ \eta^b \textstyle\sum_{s\in\mathcal{S}_b} p_s^b(y^b(e))\Big] \end{aligned}$$

where we now adopt the convention that $\mathcal{X}^b = (x_s, y_s)_{s\in\mathcal{S}_b}$ and $y^b(e) = (y_s(e_s))_{s\in\mathcal{S}_b}$.

[9] An interesting question for future work is the study of how these assignments $\mathcal{B}_s$ would come to be in real-world settings, as well as the identification of socially desirable assignments. Our results here provide evidence that it will likely be difficult to find these desirable assignments.

Letting $x^b = (x_i)_{i\in\mathcal{S}_b}$, the expected value of the payment p_s^b can now be calculated as

$$\begin{aligned} \mathrm{p}_s^b((c_s^b, a_s^b), e) &= \mathbb{E}[p_s^b(y^b(e))] \\ &= c_s^b - a_s^b\big(\sigma_s^2(e_s) + g_b(x^b_{\mathcal{S}_b\setminus\{s\}}, \delta_{x_s}, (\sigma_i^2(e_i))_{i\in\mathcal{S}_b\setminus\{s\}}\big) \end{aligned}$$

and the expected total payment s receives will be $\mathrm{p}_s((c_s, a_s), e) = \sum_{b\in\mathcal{B}_s} \mathrm{p}_s^b((c_s^b, a_s^b), e)$. With these refactored definitions, the individual rationality, non-negativity and incentive compatibility constraints on the buyers' optimization problems are still given by equations (8), (9) and (10), respectively.

Now letting $\boldsymbol{a}_s = \sum_{j\in\mathcal{B}_s} a_s^j$, it is straightforward to show that the first-order optimality condition in (16) holds, and the ensuing analysis in Section III-F pulls through. That is, one can show that each selection of parameters $a = (a^b)_{b\in\mathcal{B}}$ still induces a game between the data sources for which there is a dominant strategy equilibrium. Moreover, for each $s \in \mathcal{S}$, there exists an implicitly defined map $\mu_s\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ which returns the equilibrium level of effort e_s^* for each choice of $\boldsymbol{a}_s$. The constants $\underline{\boldsymbol{a}}_s$, $\overline{\boldsymbol{a}}_s$ and the set $\mathcal{A}_s$ can also be redefined for this setting in a natural way.

Following steps similar to those in Section III-G, the loss for aggregator b can now be written as:

$$\begin{aligned} L^b((c^b, a^b), (c^{-b}, a^{-b})) &= \textstyle\sum_{i\in\mathcal{S}_b} \gamma_i^b \sigma_i^2(\mu_i(\boldsymbol{a}_i)) \\ &+ \sum_{i\in\mathcal{S}\setminus\mathcal{S}_b} \gamma_i^b \sigma_i^2(\mu_i(\boldsymbol{a}_i)) \\ &+ \eta^b \textstyle\sum_{i\in\mathcal{S}_b} \Big(c_i^b - a_i^b \Big[\sum_{l\in\mathcal{S}_b} \xi_{i,l}^b \sigma_l^2(\mu_l(\boldsymbol{a}_l))\Big]\Big) \end{aligned} \tag{35}$$

where we define $\beta_s^b = h_b(x_s, x^b, F_b)$ and then define $\gamma_s^b = \beta_s^b - \sum_{j\in\mathcal{B}_s\setminus\{b\}} \zeta_j^b \beta_s^j$ for each $b \in \mathcal{B}$ and $s \in \mathcal{S}_b$. Note that if $s \notin \mathcal{S}_b$, then we do not need to define the constant β_s^b, since the query response that s produces does not factor into the estimator that aggregator b constructs for f. On the other hand, we do need to define γ_s^b for each $s \in \mathcal{S}$ and $b \in \mathcal{B}$, since we have assumed that s has agreed to provide at least one aggregator with the reading y_s. However, we note that the second term on the right hand side of (35) does not depend on any of aggregator b's decision variables.

Similarly, we only need to define the parameter $\xi_{i,l}^b$ if both $i, l \in \mathcal{S}_b$. In the case that data sources i and l both accept payment from aggregator b we then define:

$$\xi_{i,l}^b = \begin{cases} h_b(x_l, x^b_{-i}, \delta_{x_i}) & i \neq l \\ 1 & i = l \end{cases} \tag{36}$$

Applying the preceding analysis, aggregator b's optimization problem can now be re-written as:

$$\begin{aligned} \min_{(c^b, a^b)}\ & L^b((c^b, a^b), (c^{-b}, a^{-b})) \\ \text{s.t.}\ & \sum_{j\in\mathcal{B}_s} \left[c_s^j - a_s^j \left(\sum_{i\in\mathcal{S}_j} \xi_{s,i}^j \sigma_i^2(\mu_i(\boldsymbol{a}_i))\right)\right] \geq \mu_s(\boldsymbol{a}_s), \\ & \qquad\qquad \forall s \in \mathcal{S}_b \\ & c_s^b - a_s^b \left(\sum_{i\in\mathcal{S}_b} \xi_{s,i}^b \sigma_i^2(\mu_i(\boldsymbol{a}_i))\right) \geq 0,\ \forall s \in \mathcal{S}_b \\ & \boldsymbol{a}_s \in \mathcal{A}_s,\ \forall s \in \mathcal{S}_b \qquad a_s^b \geq 0,\ \forall s \in \mathcal{S}_b \end{aligned} \tag{37}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

where L^b is now defined by (35). Note that the optimization facing aggregator b in (37) is quite similar to the optimization in (20), save the modifications to some IR and non-negativity constraints. As we shall see in the statement of Theorem 4 and Corollary 2 below, when there are GN equilibria in the game between the buyers, the removal of some of these constraints will affect the degeneracy previously seen in the c parameters.

Before stating our generalization to Theorem 1 we redefine some final notation. First, we define: $\boldsymbol{\gamma}_s = \sum_{j\in\mathcal{B}_s}\gamma_s^j$, where we emphasize that $\boldsymbol{\gamma}_s$ does not depend on γ_s^b if $b \notin \mathcal{B}_s$, since this term will fall out when characterizing the optimality condition for aggregator b, and not affect the existence of GN equilibria. We then define for each $s \in \mathcal{S}$ and $b \in \mathcal{B}$: $q_s^b(a) = a_s^b\left(\sum_{i\in\mathcal{S}_b}\xi_{s,i}^j\sigma_i^2(\mu_i(\boldsymbol{a}_i))\right)$ and $q_s(a) = \sum_{j\in\mathcal{B}_s}q_s^j(a) + \mu_s(\boldsymbol{a}_s)$.

Theorem 4. *Consider the game $\{P_b(\cdot)\}_{b\in\mathcal{B}}$ where each aggregator's objective is to solve the optimization in* (37)*. Suppose that for each $s \in \mathcal{S}$, $\mathcal{E}_s = \mathbb{R}_{\geq 0}$, $\boldsymbol{\gamma}_s \geq \underline{\boldsymbol{a}}_s$. Further, suppose that $\gamma_s^b > 0$, for each s and b such that $s \in \mathcal{S}_b$. If there exists a GN equilibrium $(\bar{c},\bar{a})$ then the following conditions hold:*

1) *The set of GN equilibria in the game is given by*

$$\left\{(c,a) : a = \bar{a},\ \boldsymbol{c}_s = q_s(\bar{a}),\ c_s^b \geq q_s^b(\bar{a}),\ \forall s, \forall b \in \mathcal{B}_s\right\}$$

That is, the a parameters selected by the aggregators are the same across each GN equilibrium, and for each $s \in \mathcal{S}$ the equilibrium $c_s = (c_s^b)_{b\in\mathcal{B}_s}$ parameters lie in the $|\mathcal{B}_s|$-dimensional polytope defined above.

2) *The effort exerted by each data source is the same in each GN equilibrium and the efforts constitute a unique induced dominated strategy equilibrium between the data sources. More precisely, each data source exerts effort $\mu_s(\bar{\boldsymbol{a}}_s)$ in each GN equilibrium.*

The proof is almost exactly the same as the proof of Theorem 1 and is omitted here. In particular, note that the only difference we need to consider is that the aggregators now have fewer decision variables and fewer constraints on these decision parameters. The removal of these components manifests itself in the dimensionality of the polytope of equilibrium parameters.

Corollary 2. *Consider the game $\{P_b(\cdot)\}_{b\in\mathcal{B}}$ where each aggregator's objective is to solve the optimization in* (37)*, and suppose the assumptions of Theorem 4 hold. If there exists a GN equilibrium solution $(\bar{c},\bar{a})$ then the following two statements are true:*

1) *If $|\mathcal{B}_s| = 1$ for all $s \in \mathcal{S}$ then $(\bar{c},\bar{a})$ is the only GN equilibrium.*
2) *If there exists $s \in \mathcal{S}$ such that $|\mathcal{B}_s| \geq 2$ then there are an infinite number of GN equilibria.*

We state the previous result to emphasize that when even a single data source accepts incentives from more than a single aggregator an infinity of GN equilibria arise in the data market (given that any GN equilibria exist). Returning to Theorem 4, we see complete degeneracy in the equilibrium c parameters offered to any data source who sells data to multiple buyers.

An analogous generalization to Theorem 2 is also straightforward to obtain for the more general case we consider in this setting, though we omit it in the interest of brevity. The analysis conducted in the proof of Theorem 3 also follows through in a natural way. In particular, we still observe that the first order optimality conditions for the aggregators will coincide with the socially efficient choice of pricing parameters if and only if each of the non-diagonal ξ parameters is zero. In particular, this means that if $|\mathcal{B}_s| = 1$ for each $s \in \mathcal{S}$ the data market will achieve a socially efficient outcome, with regards to the exogenous assignment of data sources we have assumed has already occurred.

VI. Closing remarks

We analyzed the strategic interactions between multiple data aggregators who share a pool of data sources. Previous work showed that a single data aggregator can find unique solutions that achieve socially efficient outcomes, but we demonstrate that the same mechanisms will break down as soon as a second data aggregator enters the market. In particular, we show that there are either no GN equilibria or infinitely many, and these solutions are frequently socially inefficient. This highlights the need for further research into mechanisms for data markets when there are multiple purchasers. In particular, there is a need for mechanisms that can simultaneously handle moral hazard *and* the non-rivalrous nature of data.

Appendix

A. GNE Existence Result

Theorem 5 (Existence of GN equilibria [26])**.** *Consider a GN equilibrium problem $\{P^b(\cdot)\}_{b\in\mathcal{B}}$. Suppose that for each $b \in \mathcal{B}$, the following hold: i) the correspondence $\mathcal{M}^b\colon \prod_{i\in-b}\mathcal{Z}_i \to \mathcal{Z}_b$ is upper semi-continuous. ii) the map L^b is continuous on the graph of $\mathcal{M}^b$. iii) the map $z^{-b} \mapsto \min_{z^b\in\mathcal{M}^b(z^{-b})} L^b(z^b, z^{-b})$ is continuous. iv) for all z^{-b}, the best response set, $BR_b(z^{-b}) = \arg\min\{L^b(z^b, z^{-b}) : z^b \in \mathcal{M}^b(z^{-b})\}$ is contractable.*

Then, there exists a GN equilibrium.

B. Price of Anarchy with Bounded Effort Spaces

Theorem 6. *Suppose $\gamma_i^j > 0$, $\forall i \in \mathcal{S}$, $\forall j \in \mathcal{B}$. Further suppose that $\mathcal{E}_s = [0, e_s^{max}]$ and that $\boldsymbol{\gamma}_s \in [\underline{\boldsymbol{a}}_s, \overline{\boldsymbol{a}}_s)$ for each $s \in \mathcal{S}$. Then, there exists a GN equilibrium to the game $\{P^b(\cdot)\}_{b\in\mathcal{B}}$ between the aggregators for which the price of anarchy is equal to one if and only if for each $j \in \mathcal{B}$ and each $i, l \in \mathcal{S}$ such that $i \neq l$ we have that $\xi_{i,l}^j = 0$.*

Proof. In the case where $\mathcal{A}_s$ is bounded $\forall s \in \mathcal{S}$, we make an analogous argument as was made in Theorem 3. Suppose that $\boldsymbol{a}_s = \boldsymbol{\gamma}_s, \forall s \in \mathcal{S}$ as is needed for the socially optimal solution. Then, by assumption, for each $s \in \mathcal{S}$, we have $\underline{\boldsymbol{a}}_s < \boldsymbol{a}_s < \overline{\boldsymbol{a}}_s$, and thus, it must be true that in equilibrium the a–parameters are given by an equation of the form $a = \Xi a + \gamma$, since the best response for each aggregator is given by $a_s^b = \gamma_s^b + \sum_{j\in-b}\sum_{l\in-s} a_l^j\xi_{l,s}^j$, which is the second option in (28). However, again it cannot be the case that $\gamma_s = \boldsymbol{a}_s$ for each $s \in \mathcal{S}$ if $\xi_{i,l}^j \neq$ for some $j \in \mathcal{B}$ and $i, l \in \mathcal{S}$ such that $i \neq l$. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2922190, IEEE Transactions on Automatic Control

REFERENCES

[1] M. Babaioff, R. Kleinberg, and R. Paes Leme, "Optimal mechanisms for selling information," in *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 2012, pp. 92–109.

[2] D. Bergemann, A. Bonatti, and A. Smolin, "The design and price of information," *American Economic Review*, vol. 108, no. 1, pp. 1–48, 2018.

[3] D. G. Dobakhshari, P. Naghizadeh, M. Liu, and V. Gupta, "A reputation-based contract for repeated crowdsensing with costly verification," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 5243–5248.

[4] D. Bergemann and A. Bonatti, "Selling cookies," *American Economic Journal: Microeconomics*, vol. 7, no. 3, pp. 259–94, 2015.

[5] I. Caragiannis, A. Procaccia, and N. Shah, "Truthful univariate estimators," in *International Conference on Machine Learning*, 2016, pp. 127–135.

[6] Y. Chen, N. Immorlica, B. Lucier, V. Syrgkanis, and J. Ziani, "Optimal data acquisition for statistical estimation," in *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 2018, pp. 27–44.

[7] N. Miller, P. Resnick, and R. Zeckhauser, "Eliciting informative feedback: The peer-prediction method," *Management Science*, vol. 51, no. 9, pp. 1359–1373, 2005.

[8] D. Prelec, "A bayesian truth serum for subjective data," *science*, vol. 306, no. 5695, pp. 462–466, 2004.

[9] B. Faltings, J. J. Li, and R. Jurca, "Incentive mechanisms for community sensing," *IEEE Transactions on Computers*, vol. 63, no. 1, pp. 115–128, 2014.

[10] A. Dasgupta and A. Ghosh, "Crowdsourced judgement elicitation with endogenous proficiency," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 319–330.

[11] G. Radanovic and B. Faltings, "Incentive schemes for participatory sensing," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 1081–1089.

[12] V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes, "Informed truthfulness in multi-task peer prediction," in *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 2016, pp. 179–196.

[13] Y. Liu and Y. Chen, "Machine-learning aided peer prediction," in *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 2017, pp. 63–80.

[14] W. Vickrey, "Counterspeculation, auctions, competitive sealed tenders," *J. Finance*, vol. 16, no. 1, pp. 8–37, 1961.

[15] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[16] S. Chawla, J. D. Hartline, and B. Sivan, "Optimal crowdsourcing contests," *Games and Economic Behavior*, 2015.

[17] Y. Cai, C. Daskalakis, and C. Papdimitriou, "Optimum statistical estimation with strategic data sources," in *JMLR: Workship and Conf. Proc.*, vol. 40, 2015, pp. 1–17.

[18] F. Farokhi, I. Shames, and M. Cantoni, "Budget-constrained contract design for effort-averse sensors in averaging based estimation," *arXiv preprint arXiv:1509.08193*, 2015.

[19] ——, "Promoting truthful behavior in participatory-sensing mechanisms," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1538–1542, Oct 2015.

[20] D. G. Dobakhshari, N. Li, and V. Gupta, "An incentive-based approach to distributed estimation with strategic sensors," in *Proc. 55th Conf. Decision Control*. IEEE, 2016, pp. 6141–6146.

[21] F. Farokhi, A. M. H. Teixeira, and C. Langbort, "Estimation with strategic sensors," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 724–739, Feb 2017.

[22] T. Westenbroek, R. Dong, L. J. Ratliff, and S. S. Sastry, "Statistical estimation with strategic data sources in competitive settings," in *Proc. 56th IEEE Conf. Decision and Control*, 2017.

[23] F. Facchinei and C. Kanzow, "Generalized nash equilibrium problems," *4or*, vol. 5, no. 3, pp. 173–210, 2007.

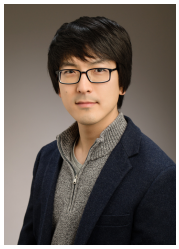
[24] P. Bolton and M. Dewatripont, *Contract theory*. MIT press, 2005.

[25] S. Stańczak, M. Wiczanowski, and H. Boche, *Chapter 2: On the Positive Solution to a Linear System with Nonnegative Coefficients*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 51–68.

[26] S. Park, "Existence theorems for generalized nash equilibrium problems," *Banged Int. J. Math & Comp. Sci.*, vol. 1, no. 1, pp. 42–51, apr 2015.

[27] T. Roughgarden, "Routing games," in *Algorithmic game theory*. Cambridge University Press, 2007, ch. 18, pp. 461–486.

Tyler Westenbroek is a graduate student at the University of California, Berkeley, pursuing a Ph.D. in Electrical Engineering and Computer Science. He graduated, with top honors, from Washington University in Saint Louis, receiving a B.S. (2016) with majors in Systems Engineering and Computer Science. His interests lie primarily in the areas of Hybrid Systems, Optimal Control, and Optimization.



Roy Dong is a Research Assistant Professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He was a postdoctoral researcher and visiting lecturer at University of California, Berkeley from 2017 to 2018, where he also received his Ph.D. in Electrical Engineering and Computer Sciences in 2017. Prior to his graduate studies, he received a B.S. Honors in Economics and a B.S. Honors in Computer Engineering from Michigan State University in 2010. He is the recipient of the National Science Foundation Graduate Research Fellowship (2011).

Lillian J. Ratliff (S'08–M'15) is an Assistant Professor in Electrical Engineering (EE) at the University of Washington, Seattle. Prior to joining UW she was a Postdoctoral Researcher in Electrical Engineering and Computer Sciences at the University of California, Berkeley where she also obtained her Ph.D. in 2015. She obtained a B.S. in Mathematics (2008) and a B.S. (2008) and M.S. (2010) in EE all from the University of Nevada, Las Vegas. Her research interests lie at the intersection of game theory, optimization, and learning. She is the recipient of the National Science Foundation Graduate Research Fellowship (2009) and the CISE Research Initiation Initiative Award (2017).

S. Shankar Sastry (S'79–M'80–SM'95–F'95) is with the University of California, Berkeley where he is faculty director of the Blum Center for Developing Economies. He served as the Dean of Berkeley's College of Engineering from 2007-2018. He received his B. Tech from the Indian Institute of Technology, and M.S., M.A. (Math), Ph.D. in Engineering from Berkeley. He has served on the faculties of MIT and Harvard. His areas of personal research are design of resilient network control systems, autonomous systems, computer vision, nonlinear and adaptive control, and hybrid and embedded systems. He is also a member of the National Academy of Engineering and the American Academy of Arts and Sciences. He received an honorary M.A. from Harvard and honorary doctorates from the Royal Swedish Institute of Technology and the University of Waterloo. He has been a member of the Air Force Scientific Advisory Board and the Defense Science Board, among other national boards. He is a member of the UN Secretary General's Scientific Advisory Board. He has coauthored over 550 technical papers and 9 books. He has supervised over 85 doctoral students and over 50 MS students. His students now occupy leadership roles in several places and on the faculties of many major universities in the United States and abroad.