FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Competitive Outcomes in the Data Market: Statistical Estimation in a Strategic Setting

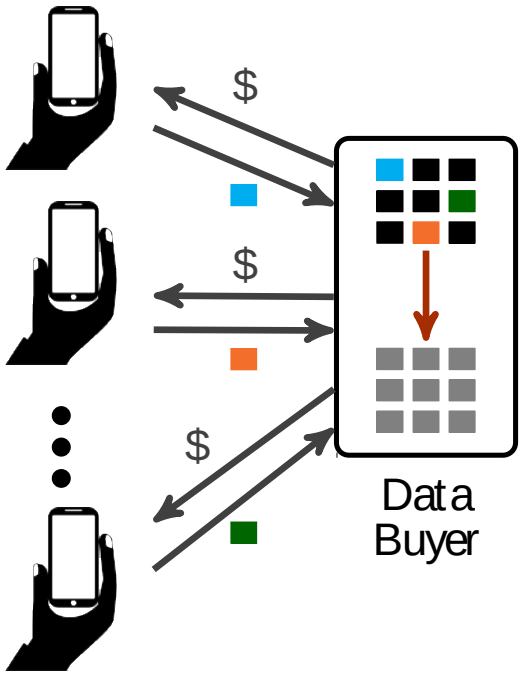
Tyler Westenbroek

with Roy Dong, Lillian Ratliff, and Shankar Sastry

Berkeley UNIVERSITY OF CALIFORNIA
Massachusetts Institute of Technology
NSF
MICHIGAN
VANDERBILT UNIVERSITY

Problem Setting

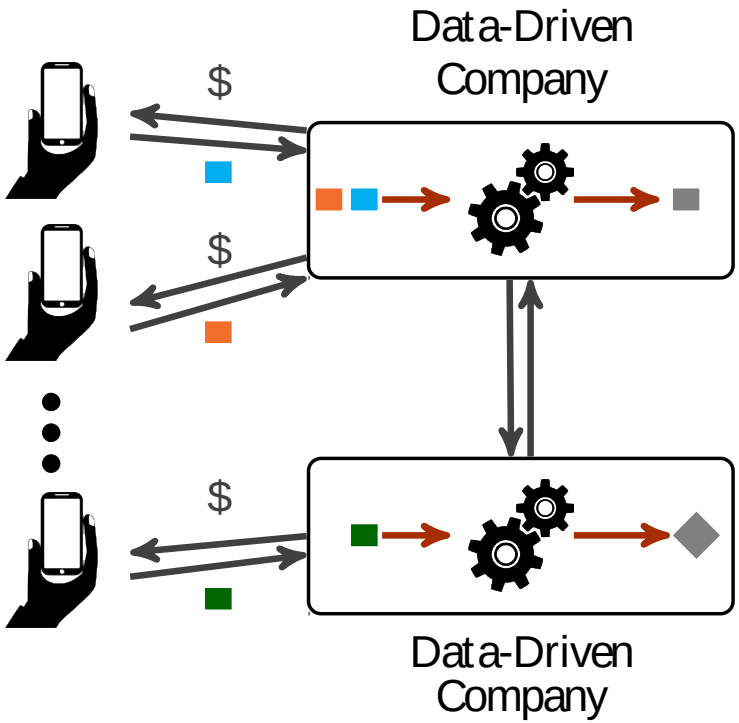
* Consider the scenario where a single central **data buyer** seeks to pool data collected from a number of **data sources** to perform some estimation task
* Recently, a class of incentive mechanisms has been studied which enable the data buyer to ensure the data reported by the data sources are of high quality, even if the sources have **incentive to obfuscate their data**



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Introducing the Data Market

* We develop a game theoretic framework to study how these mechanisms perform when there are multiple, **competing** data buyers
  * We introduce a model for the **Data Market**
  * Data buyers now also consider the value of the data that is flowing to their competitors



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Data Sources

$$y_i = f(x_i) + \xi_i$$

* The effort exerted by the data source i determines the distribution of ξ_i.
  * The effort is **not observed** by anyone other than i.
* We assume that $\mathbb{E}\xi_i = 0$ for any effort level.
* Data source i has an effort-variance function $\sigma_i: \mathbb{R} \to \mathbb{R}_+$.
  * If i exerts effort e_i, then $\mathbb{E}\xi_i^2 = \sigma_i^2(e_i)$.
  * This effort function σ_i is common knowledge to all sources and buyers.
  * Assumed to be strictly decreasing and convex.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Data Sources

* Data source i receives an incentive p_i in exchange for sharing their data (x_i, y_i).
  * Recall the effort e_i is not known to the data buyer.

* Their payoff if they choose to share their data (opt-in):

$$\mathbb{E}p_i - e_i$$

* If they choose to opt-out: 0.

Data Sources

$$\mathbb{E}p_i - e_i$$

* Assumptions in this formula:
  * Data sources are **risk-neutral**.
  * Data sources are **effort-averse**.
  * Data sources must decide whether to opt-in **ex-ante**.
  * The effort e_i can be normalized to be **comparable** with the payment p_i.

Data Buyers

* Suppose there is only one data buyer.
* The data buyer picks an **estimator** $\hat{f}$ for f.
  * This estimator depends on the data:
    $$(\vec{x}, \vec{y}) = \left((x_i)_{i \in \mathcal{S}}, (y_i)_{i \in \mathcal{S}}\right)$$
  * This estimator does **not directly know**:
    * The efforts exerted by the data sources: $\vec{e} = (e_i)_{i \in \mathcal{S}}$.
    * The variance of the reported data: $\left(\sigma_i^2(e_i)\right)_{i \in \mathcal{S}}$.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Data Buyers

* The data buyer's **optimization:**

$$\mathbb{E}\left[\left(\hat{f}_{(\vec{x},\vec{y}(\vec{e}))}(X) - f(X)\right)^2 + \sum_{i \in S} p_i\right]$$

* **Constraints:**
  * The efforts are compatible with the game:
    * $\vec{e}$ is a unique dominant strategy equilibrium.
  * Individual rationality:
    * Each data source i will choose to opt-in, i.e. $\mathbb{E}p_i - e_i \geq 0$.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Structure of Payments

* The data buyer issues incentives of the form:

$$p_i(\vec{x}, \vec{y}) = c_i - d_i \left(y_i - \hat{f}_{(\vec{x}_{-i}, \vec{y}_{-i})}(x_i) \right)^2$$

Theorem [Cai, Daskalakis, Papadimitriou 2015]

* Incentives of this form can induce a unique dominant strategy equilibrium that is **individually rational.**
* Furthermore, this equilibrium can be made **incentive compatible:**

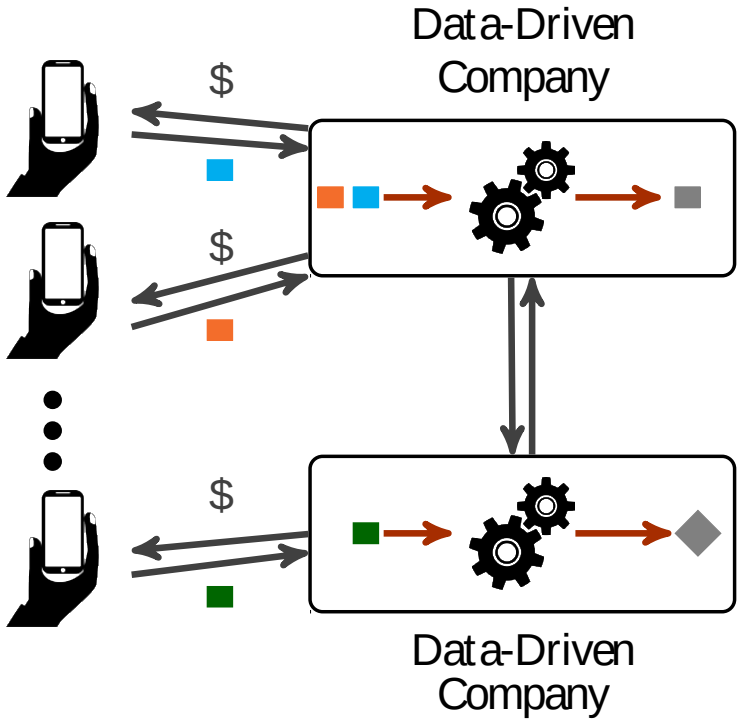
$$\mathbb{E} p_i = e_i$$

* In this case, these incentives achieve the **social optimum.**

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Multiple Data Aggregators

* Suppose there are multiple data buyers.
  * Let the set of buyers be $\mathcal{B} = \{1,2, \dots, M\}$.
* What if there is **competition** between data buyers?



[1] Cai, Daskalakis, Papadimitriou (2015)

FORCES FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Multiple Data Buyers

* Data source $i \in \mathcal{S}$ gives data buyer $j \in \mathcal{B}$ the data $\left(x_i, y_i^j\right)$.
* In return, i receives payment $p_i^j\left(\vec{x}, \overrightarrow{y^j}\right)$ from j.

* The total payments data source i receives:

$$p_i = \sum_{j \in \mathcal{B}} p_i^j\left(\vec{x}, \overrightarrow{y^j}\right)$$

Results

* ***Proposition*** [Westenbroek, Dong, Ratliff, Sastry: CDC 2017 (To Appear)]
* Let:

$$p_i = \sum_{j \in \mathcal{B}} \left[c_i^j - d_i^j \left(y_i - \widehat{f^j}_{(\vec{x}_{-i}, \vec{y}_{-i})}(x_i) \right)^2 \right]$$

* These incentives still induce a game between data sources that has a unique dominant strategy equilibrium.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Results

$$p_i^j(\vec{x},\vec{y}) = c_i^j - d_i^j\left(y_i^j - \widehat{f^j}_{\left(\vec{x}_{-i},\overrightarrow{y^j}_{-i}\right)}(x_i)\right)^2$$

Proposition [Westenbroek, Dong, Ratliff, Sastry: CDC 2017 (To Appear)]

* Under incentives of this form, data sources will share the same data with all data buyers, i.e. $y_i = y_i^j$ for all j.

Game Between Data Buyers

* However, in the case of multiple data buyers, there is also a game between data buyers.

* Fix a buyer $j \in \mathcal{B}$. Their cost is:

$$J^j\left(\overrightarrow{p^j}, \overrightarrow{p^{-j}}\right) = \mathbb{E}\left[\left(\widehat{f^j}(X_j) - f(X_j)\right)^2\right] - \mathbb{E}\left[\sum_{k \in \mathcal{B}\setminus\{j\}} \delta_{j,k}\left(\widehat{f^k}(X_{j,k}) - f(X_{j,k})\right)^2\right] + \mathbb{E}\sum_{k\in\mathcal{B}} p_i^k$$

Game Between Data Buyers

$$J^j\left(\overrightarrow{p^j}, \overrightarrow{p^{-j}}\right) = \mathbb{E}\left[\left(\widehat{f^j}(X_j) - f(X_j)\right)^2\right] - \mathbb{E}\left[\sum_{k\in\mathcal{B}\setminus\{j\}} \delta_{j,k}\left(\widehat{f^k}(X_{j,k}) - f(X_{j,k})\right)^2\right] + \mathbb{E}\sum_{k\in\mathcal{B}} p_i^k$$

* **Constraints:**
  * Efforts are compatible with the induced game:

$$e_i^* = \text{argmax}_{e_i}\, \mathbb{E}\sum_{k\in\mathcal{B}} p_i^k - e_i \text{ for all } i \in \mathcal{S}$$

  * Coefficients make sense:

$$d_i^j \geq 0 \text{ for all } i \in \mathcal{S}$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Game Between Data Buyers

* Summarize these constraints for buyer j with the set $\mathcal{M}^j\left(\overrightarrow{p^{-j}}\right)$.
  * Note that the constraint set depends on the actions of the other buyers.

Definition: The **best response set** is given by:

$$BR\left(\overrightarrow{p^{-j}}\right) = \operatorname{argmin}_{\overrightarrow{p^j} \in \mathcal{M}^j\left(\overrightarrow{p^{-j}}\right)} J^j\left(\overrightarrow{p^j}, \overrightarrow{p^{-j}}\right)$$

Generalized Nash Equilibrium

Definition: A **generalized Nash equilibrium** (GNE) for the game is a vector of payments $\left(\overrightarrow{p^j}\right)_{j\in\mathcal{B}}$ such that, for all $j \in \mathcal{B}$:

$$\overrightarrow{p^j} \in BR\left(\overrightarrow{p^{-j}}\right)$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Results

* We are able to provide **necessary** and **sufficient** conditions for the existence of GNE solutions to the game between buyers

Theorem: [Westenbroek, Dong, Ratliff, Sastry: Transaction on Automatic Control (In Preparation)]

Suppose that the game between the aggregators admits a GNE solution. Then there are an infinite number of GNE solutions to the game. Moreover, there is complete ambiguity in what portion of the payments to the data source each buyer will provide.

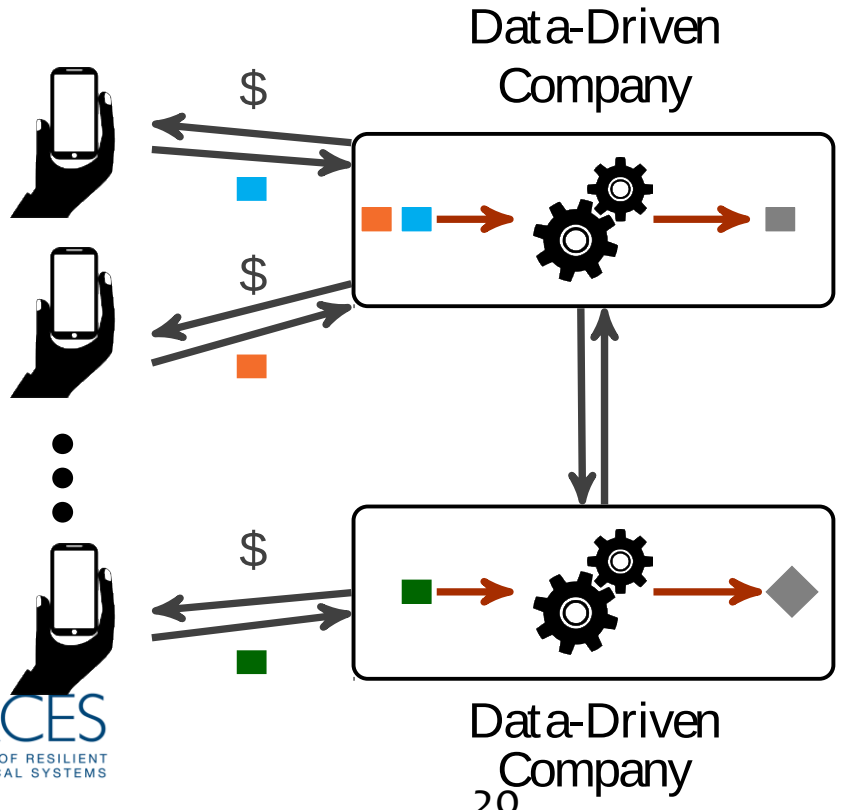
FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Characteristics of GNE Solutions

* Potential for **freeriding**
* The effort exerted by data sources may be different across different equilibria
  * Aggregators can't predict the quality of data they will receive as they could in the single aggregator case
* In most practical situations the GNE solutions are **not socially efficient**
  * One can view crowdsourcing as a **public good**, which aggregators have incentive to **over consume**

Summary of Results

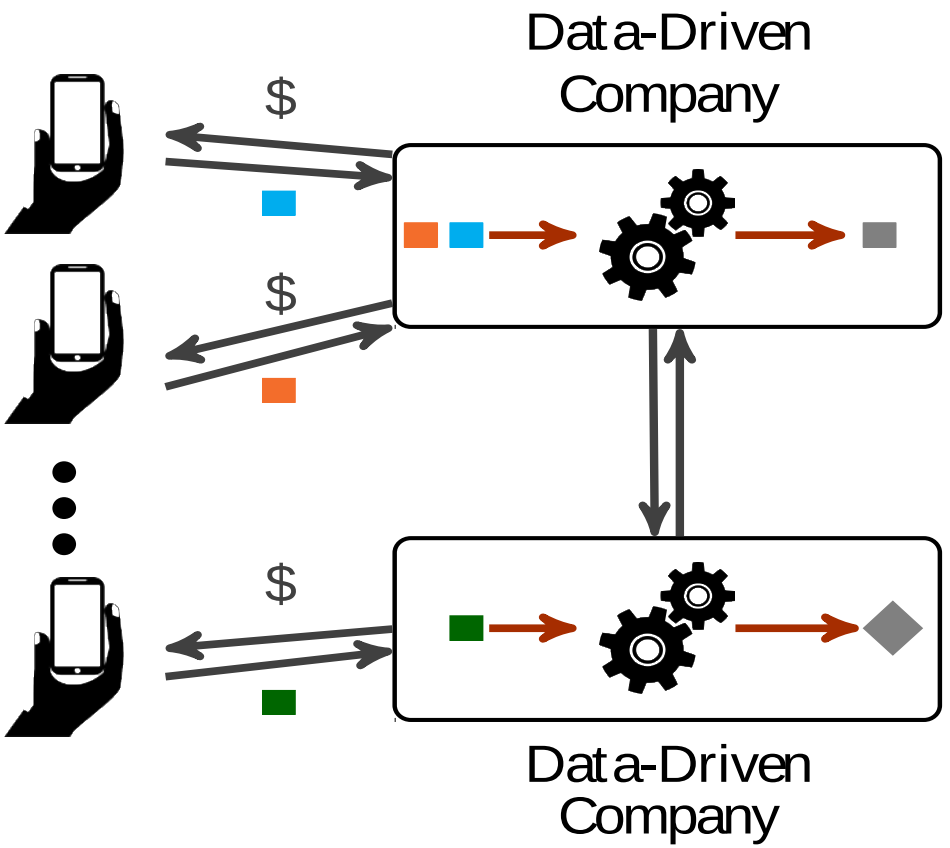
* We developed a model for the **Data Market**, and used game theoretic techniques to analyze equilibrium outcomes for a given incentive mechanism
* We demonstrated that when this mechanism was used a more realistic, competitive environment, it failed to produce the desirable results it did in the single buyer case



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Future Work

- In **future** work:
  - Expand our framework to a **dynamic** setting.
    - Explore how the value of data propagates through time
  - **Learning** dynamics and their convergence to GNE.
  - **Exclusivity** contracts between data buyers and data sources.
  - Extending results to **general classes** of payment structures.



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS