Harvard John A. Paulson
School of Engineering
and Applied Sciences

Equilibrium of Data Markets with Externality

Yiling Chen, Safwan Hossain

Motivation

- Machine learning models are only as good as the underlying data
  - Data diversity decreases overfitting and increases robustness
- Public datasets are limited in many domains
  - Healthcare, Finance, etc
- Model developers and data providers are not usually the same party

Creation of a marketplace for data providers to sell data to model developers

What makes data products special

- Reproducible at mass scale with zero marginal cost.
  - Can sell any number of copies to anyone.

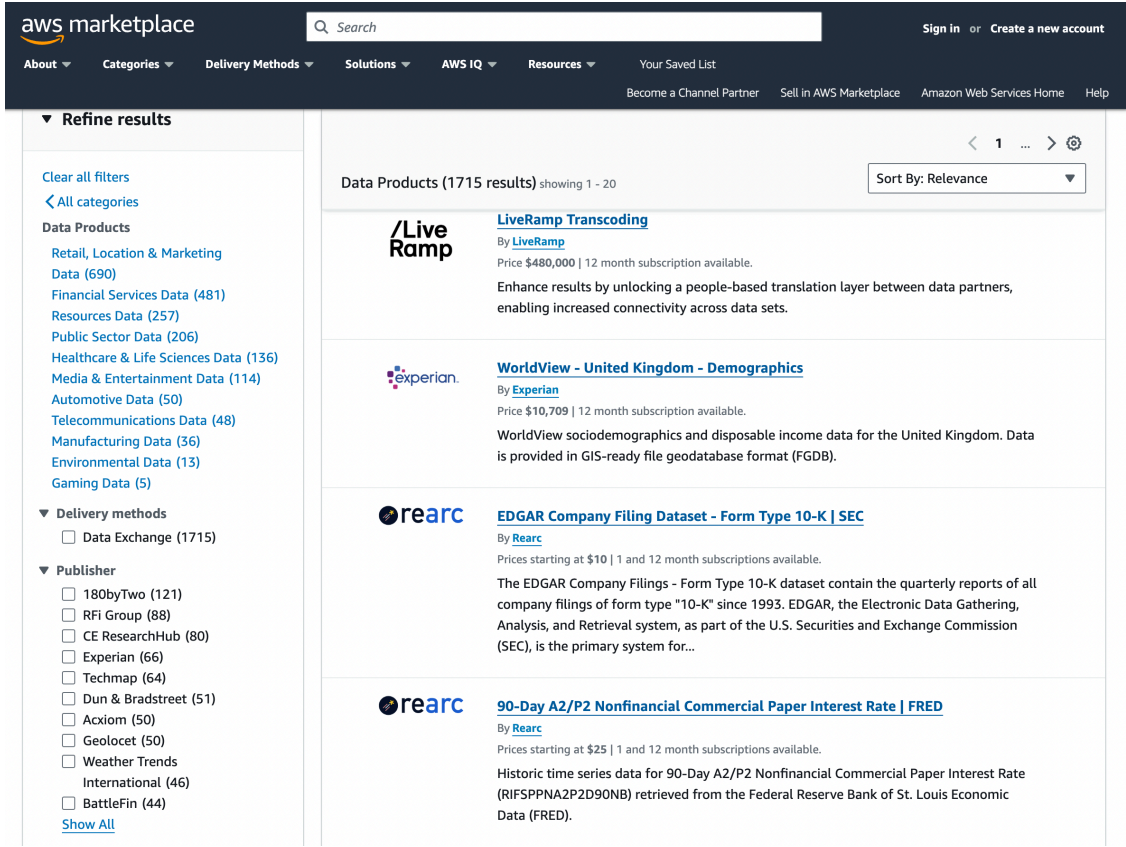
- In competitive settings, induces negative externality between buyers
  - My competitor buying high quality data can diminish my revenue.

- Hard to value a priori
  - Usefulness is only known once you have full access to it and can evaluate how it can improves your model.

- Often time-sensitive and becomes stale

Despite these unique properties, real-world data marketplaces remain quite simple.

- Sellers post fixed prices
- Buyers are not granted exclusive access
- Most are subscription based and require annual renewal.
- At best, provides a schema before buying



Our Contributions

- Model buyer interactions within such data markets as a simultaneous game
- Understand it's shortcomings and propose solutions
- Analyze it's impact under the unique characteristics of data products (unknown valuations, externality, etc)

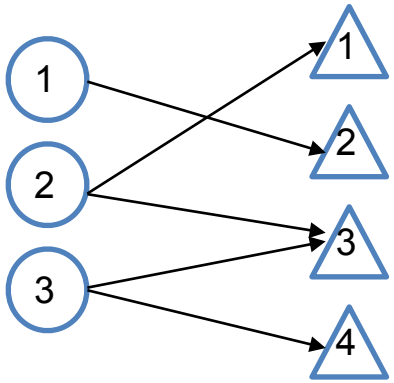
Model

- n buyers and k data providers/sellers who post fixed price
  - Buyer i can buy from any set of sellers - $\gamma_i \in 2^k$
  - Buyers simultaneously submit orders ; $S = (\gamma_1, \ldots, \gamma_n)$

- Buyer i receives net gain $G_i(\gamma_i)$ and suffers externality $E_{ij}(\gamma_j)$ due to the action of another buyer j.
  - Known as independent externality model - other models also considered*

- Platform can impose a cost $T_i(S)$ on each buyer based on the total order

Model - Example



$$\mathbb{E}[u_1] : g_1(s_2) - e_{12}(s_1, s_3) - e_{13}(s_3, s_4) - t(\cdot)$$

$$\mathbb{E}[u_2] : g_2(s_1, s_3) - e_{21}(s_2) - e_{23}(s_3, s_4) - t(\cdot)$$

$$\mathbb{E}[u_3] : g_3(s_3, s_4) - e_{31}(s_2) - e_{32}(s_1, s_3) - t(\cdot)$$

Model - Solution Concept

- Agent utilities depend on others' actions - Pure Nash equilibrium (PNE) is a natural solution concept.

- Ideally, want PNE with good welfare properties. For $S = (\gamma_1, \ldots, \gamma_n)$

$$sw(S) = \sum_{i=1} u_i(\gamma_i) \qquad S^* = \text{argmax}_S sw(S)$$

- Welfare Regret at Equilibrium (additive analogue of Price of Anarchy):

$$WRaE : sw(S^*) - \text{argmin}_{S' \in S^q} sw(S')$$

Complete Information (1)

- All buyers know the mean gains and externalities for all options
- With any constant platform cost $T_i(S) = c$:
  - PNE always exists but can have maximal WRaE
  - At equilibrium, buyers don't care about externality they impose on one another
- Platform cost should nudge agents to be cognizant of the externality they cause.
- Assume platforms have a (possibly biased) estimator these externalities - $\hat{E}_{ij}(\gamma_j)$

Complete Information (2)

- Platform charges buyers proportional to the net externality they cause:

$$T_i(S) = c + \alpha \sum_{j \neq i} \hat{E}_{ji}(\gamma_i) - \hat{E}_{ij}(\gamma_j)$$

- Cost can be negative if externality suffered is much higher than caused
  - Can be practically interpreted as discounts

- Total cost is always positive; platform does not lose money - $\sum_i T_i(S) = nc$

Complete Information (3)

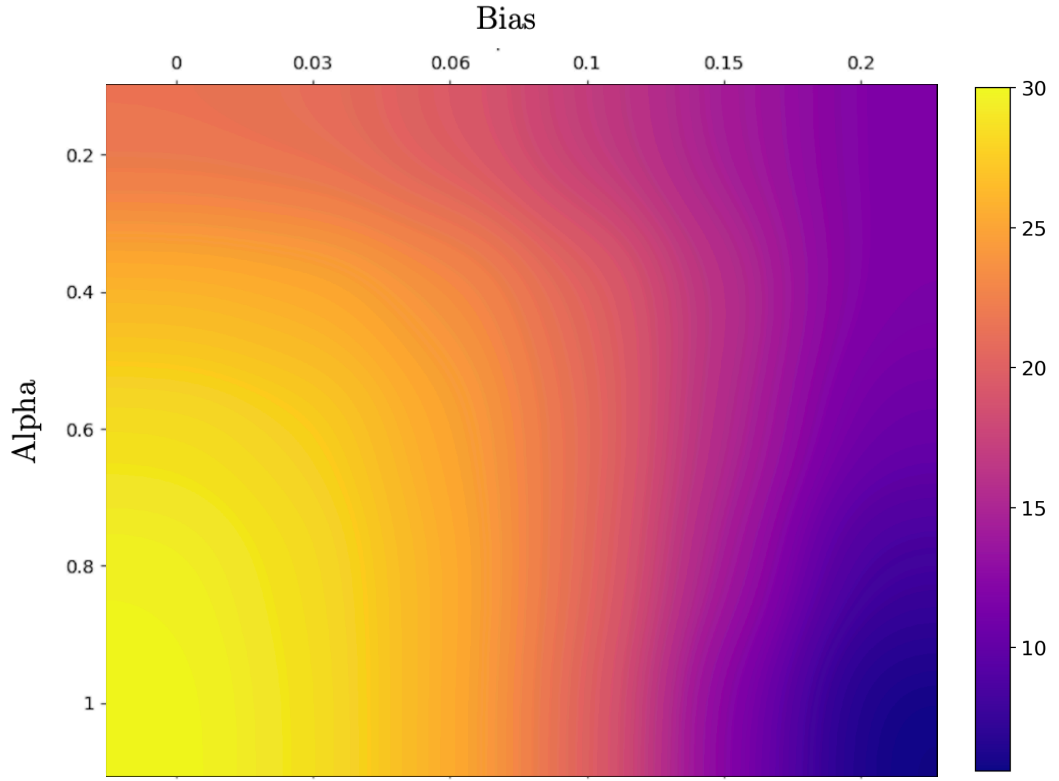
- Platform charges buyers proportional to the net externality they cause:

$$T_i(S) = c + \alpha \sum_{j \neq i} \hat{E}_{ji}(\gamma_i) - \hat{E}_{ij}(\gamma_j)$$

- A dominant strategy PNE exists under this new transaction cost.

- Then WRaE is given by $n(1-\alpha) + O(b)$
  - $O(b)$ captures the bias of platform's estimate $\hat{e}_{ij}$ of true quantity e_{ij}
  - Linearly goes to 0 as bias $\to 0$ and $\alpha \to 1$.

- Inspired by AWS marketplace
- 177 sellers across 10 categories
- Several buyers per category - each can buy from up-to 10% of sellers in their category.
- Plot increase in social welfare between constant cost equilibrium and equilibrium under proposed transaction cost.



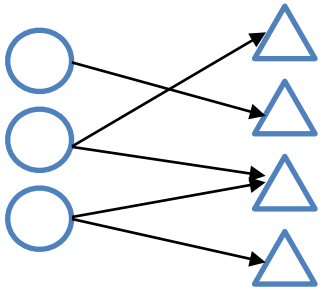
Toward a More Realistic Model

- Unknown valuations: Buyers no longer know the mean gains g_i or the associated externalities e_{ij} or $\hat{e}_{ij}$ for any choice γ.

- Repeated Interactions: Since data needs to be refreshed or its access renewed, buyers repeatedly interacting with the platform.

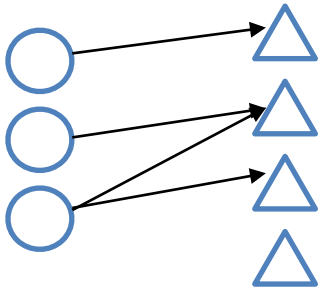
Online model where buyers make a purchase decision every time step and learn valuations through sampled realizations.

Online Model



t = 1

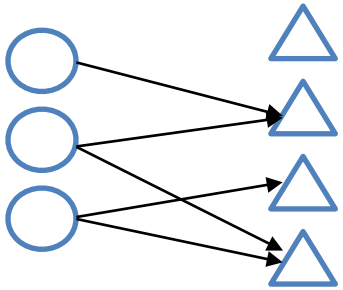
$G_i^1(\gamma), E_{ij}^1(\gamma), \hat{E}_{ij}^1(\gamma)$



t = 2

$G_i^2(\gamma), E_{ij}^2(\gamma), \hat{E}_{ij}^2(\gamma)$

• • •



t = T

$G_i^T(\gamma), E_{ij}^T(\gamma), \hat{E}_{ij}^T(\gamma)$

Online Model (2)

- Buyers face an exploration vs exploitation problem - Multi Armed Bandit
  - View γ as a binary vector of size k
  - Each set of sellers γ is an arm

- Under the proposed cost, each buyer has a dominant strategy.
  - *Reward* of an arm γ is characterized in terms of this strategy

- Problem: Exponential number of arms - worst case 2^k
  - Need to make additional assumptions about rewards structure
  - What is appropriate?

Online Model - Utility Structure

- Linear/Combinatorial Bandit:
  - Assume that utility (gain and ext) is linear: $u_i(\gamma) = w^T\gamma$
  - Utility of adding or subtracting datasets is rarely linear

- Metric Bandit:
  - Utility of "similar" arms are similar: $|u_i(\gamma_1) - u_i(\gamma_2)| \propto D_h(\gamma_1, \gamma_2)$
  - Hamming distance captures how different the purchase orders are.
  - Looser than linear and more realistic

Online Model - Metric Structure

- Hamming space is coarse metric with many ties.

- Simple ε-net style extension to UCB works poorly here.
  - Poor upper bounds on covering numbers here

- Zooming Algorithm is more flexible
  - Discrete space makes the analysis different.

Analyze the zooming algorithm in hamming space and give regret bounds for each buyer with respect to their dominant strategy.

Online Model - Individual Buyer Regert

- If $u_i(\gamma_1) - u_i(\gamma_2) \leq cD_h(\gamma_1, \gamma_2)$ - not possible to improve upon the UCB worst case.
  - Let all arms have utility within c/k of each other.
  - Metric becomes useless - get $\tilde{O}(\sqrt{2^k T})$ regret for each buyer.
- If $u_i(\gamma_1) - u_i(\gamma_2) \in [c_1 D_h(\gamma_1, \gamma_2) \pm c_2]$, can improve to $\tilde{O}(k\sqrt{kT} + 2^{0.58k})$
- Exponential dependence on k improves as utilities become more correlated. Linear bandits can be seen as an extreme end of this.

Online Model - Social Welfare Regret

- Given regret bounds for each buyer wrt their dominant strategy
- What is the corresponding regret with respect to social welfare?
  - Disentangled into regret due to learning dom strategy and offline WRaE

$$n(1-\alpha) + O(b) + \sum_i R_i^d$$

- If α is dynamic, it can be small in early rounds and increase over time as buyers have a better sense of valuations.
- In practice, buyers may have a natural shortlist of $k' < k$ sellers they consider. So regret in practice may be much better.

Richer Externality Model

- Give an online and offline characterization of a real-world data market model under a standard notion of externality.
  - Externality suffered by i due to j's action depends on this action - $e_{ij}(\gamma_j)$.

- In competitive settings, another externality model may be relevant.
  - Externality suffered by i depends on both actions - $e_{ij}(\gamma_i, \gamma_j)$

What is data market equilibrium under this joint externality model?
What is the effect of our proposed transaction cost?

Richer Externality Model - Without Constant Cost

- ϵ pure equilibrium - No player can benefit by more than ϵ by unilaterally deviating.
- With a constant transaction cost, $T_i(S) = c$, even an ϵ equilibrium may not exist for any $\epsilon < 1$.
- In instances where pure equilibrium does exist, WRaE can be maximal - n.
- Can our proposed transaction cost improve upon this?

Richer Externality Model - With Proposed Cost

- Platform charges buyers proportional to the net externality they cause:

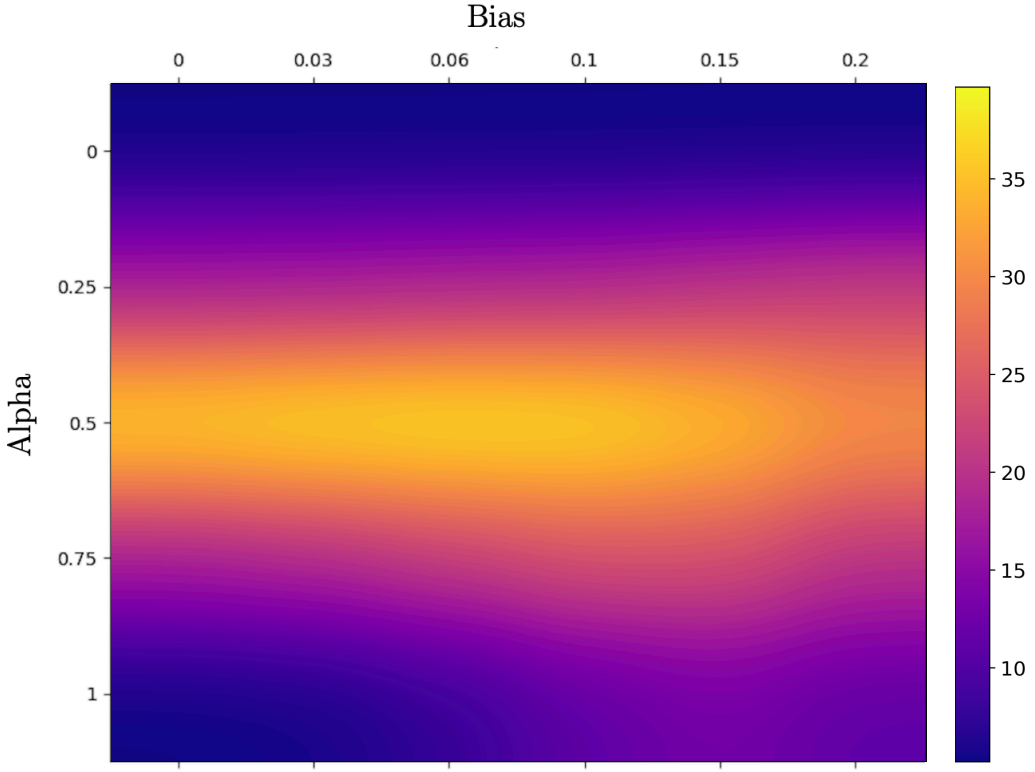
$$T_i(S) = c + \alpha \sum_{j \neq i} \hat{E}_{ji}(\gamma_i, \gamma_j) - \hat{E}_{ij}(\gamma_i, \gamma_j)$$

- An ϵ pure equilibrium always exists, with ϵ given by:

$$2\,|\,\alpha - 0.5\,|\sum_{i \neq j} \hat{e}_{ij}(\gamma_i, \gamma_j) - \hat{e}_{ji}(\gamma_i, \gamma_j) + O(b)$$

- As the externalities become symmetric and $\alpha \to 0.5$, equilibrium is exact.
- WRaE of this ϵ equilibrium is at most $n/2$.

- Same setup as before.
- Since baseline constant cost may not have any reasonable equilibrium, comparison is against myopic decision to maximize gain.



Future Directions

- Online analysis for the joint externality model.
- Elicitation approaches toward estimating externality.
- Formally define and study the space of “simple” transaction costs.
- How are sellers affected by the equilibrium of these cost structures.
    - Incorporating the strategic perspectives of sellers overall.

Thank you!

Harvard John A. Paulsor
School of Engineering
and Applied Sciences