Too Much Data: Prices and Inefficiencies in Data Markets

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, Asu Ozdaglar

MIT

January 2021, AEA

The Issue

- Data "transactions" are already pervasive as users receive valuable services and discounts in exchange of data from many online platforms.
- This is likely to grow in the next decade:
  - as social media applications, new mobile apps, and integrated technology such as IOT spread,
  - as demand for data multiplies with more extensive use of AI and machine learning techniques, and
  - as explicit data prices become widespread.
- Are we becoming better off (perhaps hugely better off) because of increasing data transactions?
- Or is there a downside?

Why Theory?

- This may appear to be an empirical question.
- But we may need a better conceptual framework for thinking about this problem.
- Existing studies approach this question from two angles:
  - The general presumption is that data creates positive externalities (for innovation, better allocation of resources), so whatever is not internalized by markets contributes to the social value of data (Varian, 2009, Jones and Tonetti, 2019, Veldkamp et al., 2019).
  - Privacy concerns weigh against this, but don't seem to be important because individuals do not appear to value their privacy (e.g., Athey et al. (2017)), but importantly this is not what they report. . . .
- We argue that this existing conceptual framework is not sufficient and once we depart from it, there may be significant costs from data.

This Paper

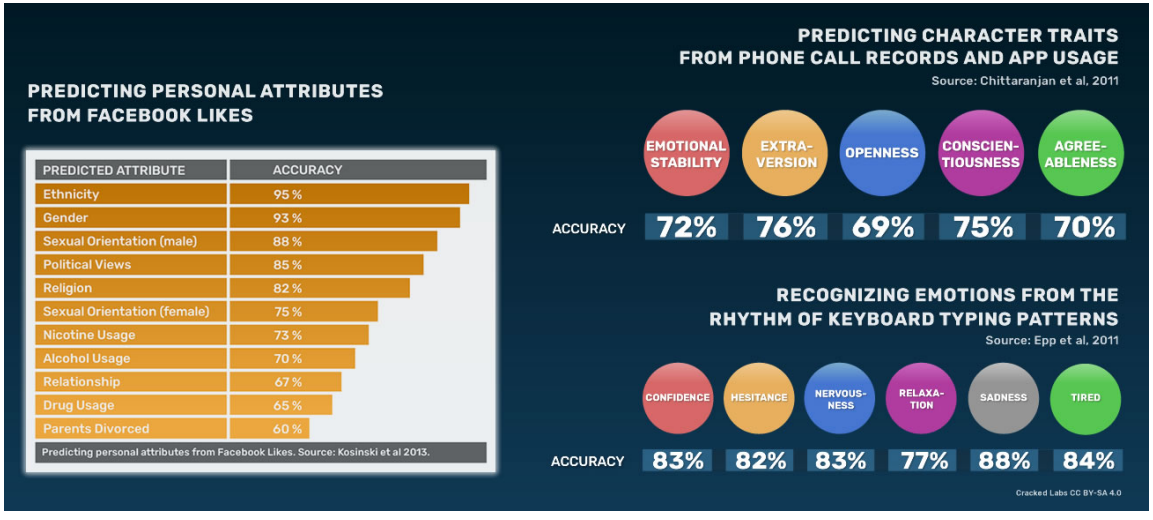
- Data sharing by one user reveals relevant data about others.
- When users value their privacy, this is a *negative externality*.
- But more importantly, because the value of information is *submodular*, these data externalities depress the price of data.
- This implies that:
  - Negative externalities from data may be substantial perhaps as large as or larger than positive externalities.
  - We cannot understand these negative externalities by just looking at the willingness of users to protect their privacy.
- We characterize data market equilibria and their efficiency properties, and provide conditions under which equilibria are inefficient and shutting down data markets improves welfare.
- We do this first with the monopoly platform, then under different types of competition and incomplete information.

The Nature of Negative Externalities I

- There are two types of negative data externalities, playing complementary roles in practice.
  1. *Direct information:*
     A post on Facebook or general online activity on other platforms will typically reveal direct information about a user's friends and acquaintances.
     An extreme example is *Cambridge Analytica*, which was able to obtain useful information from 270,000 Facebook users who downloaded the app "this is your digital life" about 50 million other users.

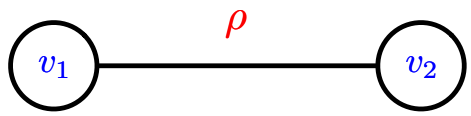
The Nature of Negative Externalities II

- There are two types of negative data externalities, playing complementary roles in practice.
  2. *Predictive information:*
     Users' behavior and data may enable platforms to predict the behavior and preferences of other users with similar demographic characteristics.
     For example, the favorite restaurants, bars and TV shows of individuals living in a location/age/gender/education/occupation cell has huge predictive power for others with the same characteristics.



An Example

- This example illustrates the main issues in our model and introduces some of the main concepts.



- Suppose there are two users, 1 and 2, each with personal data correlated with their type (assumed to be normally distributed).
- Suppose that the type of the two users is correlated with correlation coefficient ρ.
- The platform would like to predict user type and users value their privacy so they are subject to disutility v_i per unit of information leaked about them. Normalize valuation of the platform to 1.
- Suppose $v_1 < 1$. Then user 1 will always share her data and there is positive social surplus from this.

An Example (continued)

- But now also suppose $v_2 > 1$ and $\rho > 0$. Then there is a negative externality from data sharing.
- If $\rho \simeq 1$, and v_2 is sufficiently large, then data sharing by user 1 reduces (utilitarian) welfare.
- Moreover, since $\rho \simeq 1$, after user 1 shares her data, user 2 will also prefer to share her own data (even though she values her privacy a lot).
- This will lead to zero data prices even though both users might be valuing privacy by much more than this.
- We will see the same forces in our general model and also some additional strategic interactions.

Related Literature

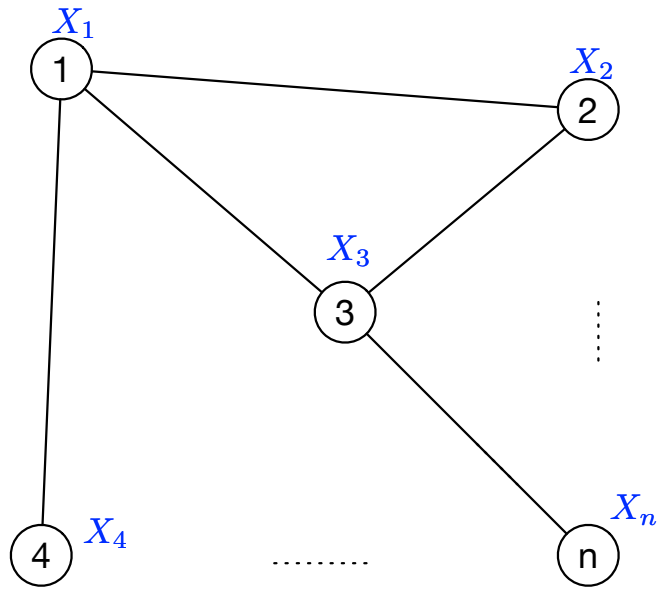
- We are related to two literatures:
  1. Privacy: [Warren and Brandeis 1890], [Westin 1968], [Posner 1981], [Varian 2009], [Goldfarb and Tucker 2012] , [Acquisti and Taylor 2016].
  2. Information markets: [Admati and Pfleiderer 1986], [Taylor 2004] , [Bergemann and Bonatti 2015], [Horner and Skrzypacz 2016], [Bergemann et al. 2018].
- Most closely related are:
  - Early papers on externalities and data sharing: [MacCarthy 2010] and [Fairfield and Engel 2015].
  - Recent work on data externalities by [Choi et al. 2019].
  - Recent work on information markets by [Bergemann et al. 2019].

Roadmap

- Introduction
- Model
- Equilibrium with a monopoly platform
- Equilibrium and inefficiency with competing platforms
- Incomplete information
- Regulation and policy
- Conclusion

Information

- x_i: type of user $i \in \mathcal{V} = \{1, \ldots, n\}$; realization of a r.v. X_i.
- $\mathbf{X} = (X_1, \ldots, X_n) \sim N(\mathbf{0}, \Sigma)$, with $\Sigma_{ii} = \sigma_i^2$.
- **S**: vector of data, where $S_i = X_i + Z_i$ for $Z_i \sim N(0, 1)$ if i shares her data.



Leaked Information and Payoffs

- $a_i \in \{0, 1\}$: data sharing action of user $i \in \mathcal{V}$, with $\mathbf{a} = (a_1, \ldots, a_n)$.
- $\mathbf{S_a}$: $= (S_i \ : \ i \in \mathcal{V} \text{ s.t. } a_i = 1)$ is platform's data.

Definition (Leaked information)

Leaked information of (or about) user $i \in \mathcal{V}$ is the reduction in the MSE of the best estimator of the type of user i:

$$\mathcal{I}_i(\mathbf{a}) = \sigma_i^2 - \min_{\hat{x}_i} \mathbb{E}\left[\left(X_i - \hat{x}_i\left(\mathbf{S_a}\right)\right)^2\right].$$

Leaked Information and Payoffs

- Payoff of Platform:

$$U(\mathbf{a}, \mathbf{p}) = \sum_{i \in \mathcal{V}} \mathcal{I}_i(\mathbf{a}) - \sum_{i \in \mathcal{V}:\ a_i = 1} p_i.$$

  - p_i: denotes payments to user i from the platform.

- Payoff of user i:

$$u_i(a_i, \mathbf{a}_{-i}, \mathbf{p}) = \begin{cases} p_i - v_i \mathcal{I}_i\left(a_i = 1, \mathbf{a}_{-i}\right), & a_i = 1 \\ \\ -v_i \mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}\right), & a_i = 0, \end{cases}$$

  - $v_i \geq 0$: user i's value of privacy.

Equilibrium Concept

Definition (User equilibrium)

Given the price vector $\mathbf{p} = (p_1, \ldots, p_n)$, an action profile $\mathbf{a}$ is user equilibrium if for all $i \in \mathcal{V}$,

$$a_i \in \operatorname{argmax}_{a \in \{0,1\}} u_i(a_i = a, \mathbf{a}_{-i}, \mathbf{p}).$$

$\mathcal{A}(\mathbf{p})$: The set of user equilibria at price $\mathbf{p}$.

Definition (Stackelberg equilibrium)

A pair $(\mathbf{p}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}})$ of price and action vectors is a pure strategy Stackelberg equilibrium if $\mathbf{a}^{\mathrm{E}} \in \mathcal{A}(\mathbf{p}^{\mathrm{E}})$ and there is no profitable deviation for the platform, i.e.,

$$U(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) \geq U(\mathbf{a}, \mathbf{p}), \quad \text{for all } \mathbf{p} \text{ and for all } \mathbf{a} \in \mathcal{A}(\mathbf{p}).$$

First Best

- Utilitarian welfare = social surplus is

$$\text{Social surplus}(\mathbf{a}) = \sum_{i \in \mathcal{V}} (1 - v_i) \mathcal{I}_i(\mathbf{a}).$$

- $\mathbf{a}^{\text{FB}}$ (First best): sharing profile that maximizes social surplus.

Proposition

The first best involves $a_i^{\text{FB}} = 1$ *if*

$$\sum_{j \in \mathcal{V}} (1 - v_j) \frac{\left(\text{Cov}\left(X_i, X_j \mid a_i = 0, \mathbf{a}_{-i}^{\text{FB}}\right)\right)^2}{1 + \sigma_j^2 - \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i}^{\text{FB}})} \geq 0,$$

and $a_i^{\text{FB}} = 0$, *otherwise.*

- When there is no correlation across individuals, then all users with $v_i < 1$ share their data and those with $v_i > 1$ do not.

Properties of Leaked Information

Lemma

1. Monotonicity: for two action profiles $\mathbf{a}$ *and* $\mathbf{a}'$ *with* $\mathbf{a} \geq \mathbf{a}'$,

$$\mathcal{I}_i(\mathbf{a}) \geq \mathcal{I}_i(\mathbf{a}'), \quad \forall i \in \{1, \ldots, n\}.$$

2. Submodularity: for two action profiles $\mathbf{a}$ *and* $\mathbf{a}'$ *with* $\mathbf{a}'_{-i} \geq \mathbf{a}_{-i}$,

$$\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) \geq \mathcal{I}_i(a_i = 1, \mathbf{a}'_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}'_{-i}).$$

- **Monotonicity:** More sharing by others leads to more leaked information.
- **Submodularity:** When more is revealed by others, there is less left for one to reveal about oneself.
- All of our results follow from monotonicity and submodularity properties and I will return to a more general versions of these results in a little bit.

User Equilibrium

Lemma

For any $\mathbf{p}$*, the set* $\mathcal{A}(\mathbf{p})$ *is a complete lattice, and thus has a least and a greatest element.*

- The (second-stage) game is supermodular because of the submodularity of the information.
- The lemma follows from Tarski's theorem.

Existence of Equilibrium

Theorem

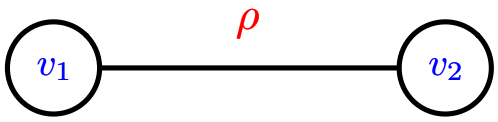
An equilibrium always exists: $\mathbf{a}^{\mathrm{E}}$ *and* $\mathbf{p}^{\mathrm{E}}$ *such that*

$$U(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) \geq U(\mathbf{a}, \mathbf{p}), \qquad \text{for all } \mathbf{p} \text{ and for all } \mathbf{a} \in \mathcal{A}(\mathbf{p}).$$

- Even if there are multiple equilibria, they all yield the same to the platform.
- This is because the platform is the Stackelberg leader.

An Illustrative Example

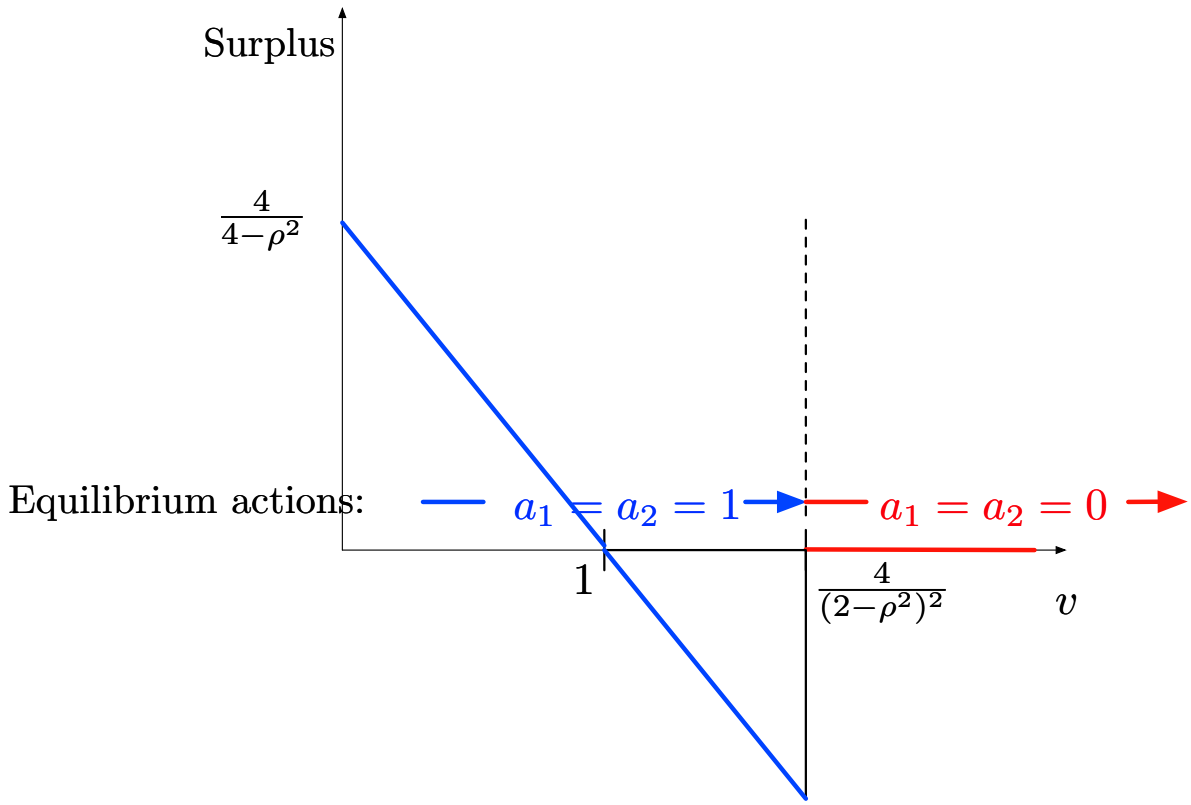
- Two users with valuations $v_1 = v_2 = v$ and correlation $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$



  - Total payment to users is non-monotone in the number of users who share:
    - Total payment to induce both users to share is $v\frac{(2-\rho^2)^2}{4-\rho^2}$.
    - Total payment to induce one user to share is $\frac{v}{2}$.
    - $\Rightarrow$ for $\rho^2 \geq \frac{7-\sqrt{17}}{4} \approx 0.71$, the platform pays less to have both users share their data.

An Illustrative Example (continued)

- Equilibrium (social) surplus is non-monotonic in the users' value of privacy:
  - $v \leq 1$: both users share and equilibrium surplus is positive.
  - $v \geq \frac{4}{(2-\rho^2)^2}$: users do not share and equilibrium surplus is zero.
  - $v \in [1, \frac{4}{(2-\rho^2)^2}]$: the platform induces sharing by both and equilibrium surplus is negative.



Equilibrium Prices

$\mathbf{p^a}$: "equilibrium price vector" — the least (element-wise minimum) price vector that sustains sharing profile $\mathbf{a}$.

Theorem

For any action profile $\mathbf{a} \in \{0,1\}^n$,

$$\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) = \frac{\left(\sigma_i^2 - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right)^2}{(\sigma_i^2 + 1) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})},$$

and $\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) = \mathbf{d}_i^T \left(I + D_i\right)^{-1} \mathbf{d}_i$, *where* D_i *is the matrix obtained by removing rows-columns* i *and rows-columns* j *for which* $a_j = 0$, *and* $\mathbf{d}_i = (\Sigma_{ij} : j \text{ s.t. } a_j = 1)$.

$$p_i^{\mathbf{a}} = \begin{cases} v_i \frac{\left(\sigma_i^2 - \mathcal{I}_i(a_i=0,\mathbf{a}_{-i})\right)^2}{(\sigma_i^2+1) - \mathcal{I}_i(a_i=0,\mathbf{a}_{-i})}, & a_i = 1, \\ 0, & a_i = 0. \end{cases}$$

- Prices are at the reservation value of users given that the platform is the Stackelberg leader.

Equilibrium Prices: Decomposition

- Covariance matrix of (X_i, S_i) conditional on $\mathbf{a}_{-i}$ and $a_i = 0$:

$$\begin{pmatrix} \operatorname{var}(X_i|a_i = 0, \mathbf{a}_{-i}) & \operatorname{cov}(X_i, S_i|a_i = 0, \mathbf{a}_{-i}) \\ \operatorname{cov}(X_i, S_i|a_i = 0, \mathbf{a}_{-i}) & \operatorname{var}(S_i|a_i = 0, \mathbf{a}_{-i}) \end{pmatrix}.$$

- Then:

$$\operatorname{var}(X_i|a_i = 1, \mathbf{a}_{-i}) = \operatorname{var}(X_i|a_i = 0, \mathbf{a}_{-i}) - \frac{\operatorname{cov}^2(X_i, S_i|a_i = 0, \mathbf{a}_{-i})}{\operatorname{var}(S_i|a_i = 0, \mathbf{a}_{-i})}.$$

- Using the definition of leaked information and $S_i = X_i + Z_i$:

$$\begin{aligned} \mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) &= \sigma_i^2 - \operatorname{var}(X_i|a_i = 1, \mathbf{a}_{-i}) \\ &= \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) + \frac{\operatorname{var}^2(X_i|a_i = 0, \mathbf{a}_{-i})}{1 + \operatorname{var}(X_i|a_i = 0, \mathbf{a}_{-i})} \\ &= \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) + \frac{(\sigma_i^2 - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}))^2}{1 + \sigma_i^2 - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})}. \end{aligned}$$

Low-Value and High-Value Users

Lemma

All users with value of privacy $v_i \leq 1$ share their data in equilibrium.

- "low-value users": $\mathcal{V}^{(l)} = \{i \in \mathcal{V} : \ v_i \leq 1\}$.
- "high-value users": $\mathcal{V}^{(h)} = \{i \in \mathcal{V} : \ v_i > 1\}$.
- $\mathbf{v}^{(h)}$ and $\mathbf{v}^{(l)}$: the vectors of valuations of privacy for high-value and low-value users, respectively.

Inefficiency

Theorem

1. *Suppose high-value users are uncorrelated with others. Then the equilibrium is efficient.*
2. *Suppose at least one high-value user is correlated with a low-value user. Then there exists* $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ *such that for* $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ *the equilibrium is inefficient.*
3. *Suppose high-value users* $\tilde{\mathcal{V}}^{(h)} \subseteq \mathcal{V}^{(h)}$ *are correlated with at least one other high-value user (and no high-low correlation). Then for each* $i \in \tilde{\mathcal{V}}^{(h)}$ *there exists* $\bar{v}_i > 1$ *such that if for any* $i \in \tilde{\mathcal{V}}^{(h)}$ $v_i < \bar{v}_i$*, the equilibrium is inefficient*

- Part 2 captures inefficiencies from externalities.
- Part 3 captures inefficiencies from depressed prices.

Are Data Markets Beneficial?

Proposition

$$\text{Social surplus}(\mathbf{a}^{\mathrm{E}}) \leq \underbrace{\sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V})}_{I} - \underbrace{\sum_{\mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)})}_{II}.$$

- Term I is an upper bound on the gain in social surplus from the sharing decisions of low-value users.
- Term II is a lower bound on the loss of privacy from high-value users.

Corollary

If

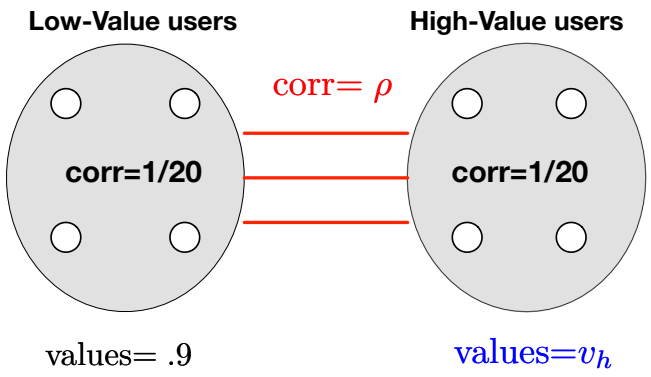
$$\sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)}) > \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}),$$

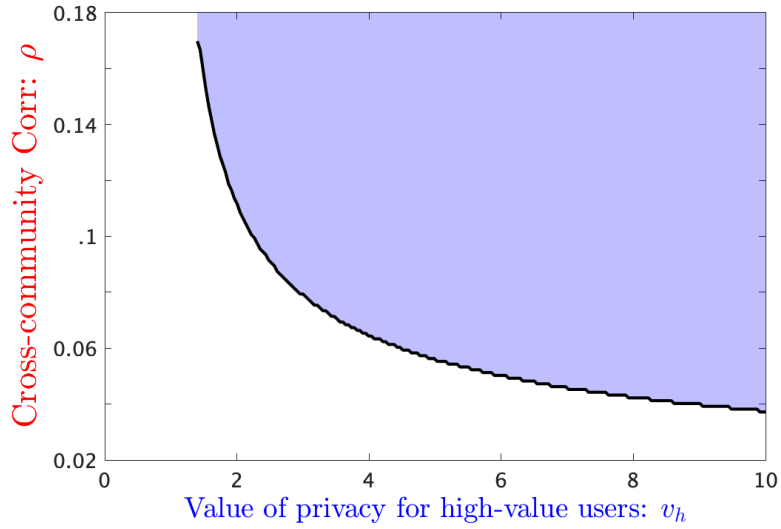
then utilitarian welfare improves when data markets are shut down.

- Straightforward to provide conditions on primitives for this to be the case.

Are Data Markets Beneficial? (continued)

- Consider a setting with two communities, each of size 10.
- The value of privacy for all users in community 1 are 0.9 and for all users in community 2 are $v_h > 1$.
- The variances of all user data are 1, within community corr are $1/20$, and the cross-community corr are ρ.
- Shaded area shows the pairs of (ρ, v_h) with negative equilibrium surplus.





Generalization

- Let us now relax the functional form restrictions in payoffs and distributions, and instead introduce the following general conditions:
  1. *No leakage with independence:* If a user i's information is independent from the information of all other users, then we have $\mathcal{I}_j(a_i = 1, \mathbf{a}_{-i}) = \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i})$.
  2. *Leakage with non-independence:* If the information of two users i and j are non-independent (given any set of other users who share), then for any action profile where user i shares her data, leaked information about user j will be non-zero.
  3. *Monotonicity*
  4. *Submodularity*
- Our baseline setup satisfies these four conditions.

Main Result

Theorem

Assume Properties 1-4 hold.

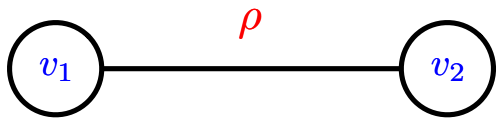
1. *Suppose every high-value user is independent from all other users. Then the equilibrium is efficient.*
2. *Suppose at least one high-value user is non-independent from a low-value user (conditional on the data shared by any set of other users). Then there exists* $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ *such that for* $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ *the equilibrium is inefficient.*
3. *Suppose high-value users* $\tilde{\mathcal{V}}^{(h)} \subseteq \mathcal{V}^{(h)}$ *are non-independent from at least one other high-value user (and no high-low dependencies). Then for each* $i \in \tilde{\mathcal{V}}^{(h)}$ *there exists* $\bar{v}_i > 1$ *such that if for any* $i \in \tilde{\mathcal{V}}^{(h)}$ $v_i < \bar{v}_i$*, the equilibrium is inefficient.*

Other Generalizations

1. The same results generalize when the platform does not know the correlation structure but has beliefs over it.
2. Similar results hold when there are competing platforms.
3. Competition does not necessarily improve efficiency, and may worsen it as the next example shows.

Does Competition Help Efficiency?

- Two users with correlation $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $v_1 < 1$, and constant joining value c.



  - *Competition improves equilibrium surplus:* If $v_2 \gg 1$:
    - Under monopoly, only user 1 shares.
    - With competition, users join different platforms and user 1 shares.
    - $\Rightarrow$ With competition equilibrium surplus improves because the data of user 1 does not leak information about user 2.
  - *Competition reduces equilibrium surplus:* If $v_2 < 1$:
    - Under monopoly, both users share.
    - With competition, users join different platforms and they both share.
    - $\Rightarrow$ With competition equilibrium surplus reduces because the platforms do not gain from the data externality.

How can we deal with data externalities? Taxation

- What can be done about inefficiency?
- Person-specific taxes can decentralize the first best (not surprisingly)

Proposition

Let $\mathbf{a}^{\mathrm{FB}}$ *denote the first best. Then personalized taxes satisfying*

$$t_i > \sum_{j \in \mathcal{V}^{(l)}} \sigma_j^2 + \sum_{j \in \mathcal{V}^{(h)}} v_j \sigma_j^2 \qquad \text{for } a_i^{\mathrm{FB}} = 0$$

$$t_i = 0 \qquad \text{for } a_i^{\mathrm{FB}} = 1,$$

implements the first-best action profile $\mathbf{a}^{\mathrm{FB}}$ *as the unique equilibrium.*

- But such taxes require a social planner to have too much information about each individual.
- Also uniform taxes do not always improve efficiency or economic surplus.

Mediated Data Sharing

- An alternative is to investigate alternative architectures of data markets.
- Here we make some preliminary advance in this direction.
- Consider the following

$$\text{“de-correlation” scheme : } \tilde{\mathbf{S}} = \Sigma^{-1}\mathbf{S}, \quad \text{for } \mathbf{S} = (S_1, \ldots, S_n)$$

- With this linear transformation of S, we have $\mathrm{Cov}(X, \tilde{S}) = I$:
  1. X_i and $\tilde{S}_{-i}$ have zero correlation
  2. X_i and $\tilde{S}_i$ are fully correlated

Mediated Data Sharing (II)

Lemma

With de-correlation, leaked information about user i is

$$\widetilde{\mathcal{I}}_i(\mathbf{a}) = \sigma_i^2 - \min_{\hat{x}_i} \mathbb{E}\left[\left(X_i - \hat{x}_i\left(\tilde{\mathbf{S}}_{\mathbf{a}}\right)\right)^2\right] = \begin{cases} 0, & a_i = 0, \\ \mathcal{I}_i(a_i, \mathbf{a}_{-i}), & a_i = 1. \end{cases}$$

- De-correlation removes the correlation between any user who does not wish to share her data and all other users, while maintaining the correlation among users sharing their data.
- With de-correlation, in contrast with (person-specific) taxes, the planner does not need to know the value of privacy of individuals or the exact correlation structure.
- Critically, this de-correlation procedure is different from anonymization of data because it does not hide information about the user sharing her data but about others who are correlated with this user.

Efficiency with De-correlation

Theorem

Let $(\tilde{\mathbf{a}}^{\mathrm{E}}, \tilde{\mathbf{p}}^{\mathrm{E}})$ *and* $(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}})$ *denote the equilibrium with and without the de-correlation scheme, respectively. Then*

$$\text{Social surplus}(\tilde{\mathbf{a}}^{\mathrm{E}}) \geq \max\left\{\text{Social surplus}(\mathbf{a}^{\mathrm{E}}), 0\right\}.$$

- Intuition. With de-correlation:
  - high-value users never contribute negative value. Hence social surplus is always nonnegative.
  - negative externalities are lessened, so social surplus always improves.
- But de-correlation does not guarantee first best.

Conclusion

- A contribution to our understanding of the effects of externalities in data markets.
- Main results:
  - Depressed data prices.
  - Potentially too much data being transacted.
  - Shutting down data markets may be socially beneficial.
  - Introducing mediated data transactions may improve welfare and in the presence of such interactions it is never optimal to shut down data markets.
- Much to be done.
- But most importantly, these results call for a different empirical strategy to investigate the value of data and the extent of privacy concerns, since these cannot be understood from revealed preference type arguments or from observed practices.