Digital gold? Pricing, inequality and participation in data markets[*]

George Charlson[†]

August 25, 2022

Abstract

I examine inequalities arising from biases brought by the incentives and externalities present in data markets, where a data collector ultimately provides an end-service which is beneficial. Agents receive different prices for their data, which is valued according to the extent that it is representative of the data of non-participating agents. The service provider estimates the characteristics of high-cost and minority groups with less accuracy, leading to these groups receiving lower quality services on average, and lower utility in equilibrium. Data privacy policies tend to reduce such inequalities but at the cost of consumer surplus, while a subsidy strategy targeted at increasing the utility of those disadvantaged by data markets increases consumer surplus but may also widen inequality.

[*]I would like to thank Sanjeev Goyal, Matt Elliott, Ozan Candogan, Ben Golub and Melika Liporace for their invaluable comments at various stages of this project, along with the participants at the Cambridge Networks Conference.

[†]Cambridge Janeway Institute, Austin Robinson Building, Sidgwick Ave, Cambridge CB3 9DD, gc556@cam.ac.uk

1 Introduction

As governments, NGOs and other organisations amass increasingly informative data on individuals, the emergence of so-called "data inequality" is becoming an inevitability. While there are many forms of data inequality, I consider inequalities which arise in data markets where agents of different groups participate to different extents, which leads to biases in collected datasets. This form of inequality results in organisations providing services and treatments to a population of agents knowing less about some groups than others.

Differences in participation in data markets are well documented. To cite just a few examples, women participate in surveys at a greater rate than men (see, for example, Otufowora et al 2021), minorities in the United States are less likely to participate in medical trials and public health initiatives than non-minorities (Milani et al, 2021), less educated people are less likely to participate in online surveys (Jang and Vorderstrasse, 2019) and older subjects are less likely to participate in trials and surveys that involve smartphone applications (Mulder and de Bruijne, 2019).

These differences in participation imply that inequalities can arise not just in the use of data once it is collected, but in the data collection process itself. Such inequalities have real-world effects, because service providers know less about some groups than others, which in turn leads to less effective policies and treatments being implemented for these groups. For example, the National Academies of Science, Engineering and Medicine (2022) found that differences in clinical trial participation is one driver for health disparities, which cost the United States $11 trillion.

In the model here, data inequalities arise endogenously from a market in which agents are paid by service provider for their data. The service provider offers a targeted service which is beneficial to the agents based on the data collected on some characteristic of interest. These characteristics are jointly normally distributed, and so i's participation results in the service provider being able to better estimate j's characteristic, even if j does not participate.

To understand the value of an agent's data to the service provider, the information structure can be characterised as a hypothetical weighted network, which captures the indirect and direct information learning an agent's characteristic provides at the margin; that is, taking into account agents who have already participated. The value of an agent's data reflects not their abstract informativeness, but moreover the extent to which they are representative of the characteristics of non-participating individuals.

An agent's representativeness can be expressed as a form of "intercentrality", a common measure found in much of network theory (see, for example, Ballester et al, 2006). Meanwhile, the price agents receive depends on both their participation cost and how informed the service provider is about their characteristic if they do not participate, which is tied to a form of their centrality, another common measure in networks.

The externalities associated with participation in the data collection process also impact the nature of the data collected. Externalities arise due to the fact that other agents' participation decisions result in the service provider learning about i's characteristic, which in turn reduces i's incentive to participate. Hence, the decision to participate by one agent crowds out the participation of other agents. This implies that if the end service is sufficiently beneficial, the service provider would, paradoxically, prefer the case where there are no data leakages; that is, no correlation between characteristics.

I then turn to the question of systematic inequalities driven by differences in participation. Individuals from groups with higher participation costs are underrepresented in the datasets collected on the data market, as incentivising them to participate is inefficiently costly for the service provider. Minority groups are similarly underrepresented, due to the fact that, holding participation rates constant, their characteristic is relatively less valuable to the platform's inference problem.

Policies, like GDPR, which reduce the extent to which intermediaries and platforms can learn about agents passively are of increasing interest to policy makers. A decorrelation policy, which reduces data leakages to zero, would be a highly effective form of such a policy. This policy is optimally equitable, in the sense that agents all receive the same payoff, but it reduces the effectiveness of the service provider's targeted policies and therefore, consumer utility.

The framework developed here lends itself naturally to a consideration of a targeted subsidy policy, aimed at increasing the utility of groups which lose out in the original data market. I show that in order to increase the utility of such groups, it may be optimal for a social planner to induce other groups to participate, which increases utility but further exacerbates inequality.

As such, our analysis of both these different types of policy highlight that intervention in data markets of the type analysed here often involves trading off generating increased surplus for the aggregate (or even specific groups) and equity concerns.

Literature and contribution

This paper contributes to two strands of literature: the literature on data markets and their externalities and the literature on algorithmic fairness.

The relevant data markets literature comes in two forms. First, there is large literature on the pricing and sale of data - some relevant references include Eso and

Szentes (2007), Babiaoff et al (2012), Bergemann et al (2018) and Eliaz et al (2019). Of particular relevance within this literature are contributions which revolve around data collection for the purpose of price discrimination, where agents are therefore harmed by the sale of their data - see, for example, De Corniere and De Nijs (2016), Gu et al (2019), Montes et al (2019), with Agrawal and Goldfarb (2018) and Bergemann and Bonatti (2019) providing comprehensive surveys of this literature.

The second branch of the data markets literature of relevance are papers which examine externalities generated by data. Most of this literature focuses on privacy, with MacCarthy (2011) and Fairfield and Engel (2015) providing analysis of the negative externalities associated with data leakages; Choi et al (2019) shows how such negative externalities can lead to inefficiencies and Ichihashi (2022) examines dynamic data collection by a platform.

Within this literature, the two most relevant papers are Ichihashi (2021) and Acemoglu et al (2022). The former examines how different consumer preferences and data properties can influence the nature of the externalities generated by data markets and hence also affect efficiency, profits and consumer surplus. The latter examines how the negative externalities associated with data leakages can reduce prices for data and increase data market participation through a "crowding-in" effect.

This paper contributes to (and is different from) the externalities in data markets literature in a number of ways. First, it shows that a tractable and intuitive way of characterising the pricing and value of data is by treating the information structure as a hypothetical network. In doing so, I provide insights into the effect of targeted policies aimed at increasing data market participation.

Second, by analysing the case where agents incur costs to participate in the data collection process, I generate new insights into real-world data markets which have this feature, specifically regarding differences in prices and participation across groups when the data collector is ultimately offering a beneficial service - very little of the above literature examines this form of service, with Ichihashi (2021) being the one exception. That paper, however, considers the case where participation is costless, with agents (at best) giving their data away for free.

Most fundamentally, the model provides an account of which agents receive relatively high quality services, hence generating predictions regarding data inequalities which arise from data collection that are not present in the current literature.

This last feature of the paper also implies there is a link between it and the algorithmic fairness literature, which seeks to analyse and correct inequalities generated by machine learning techniques - see Kleinberg et al (2018) and Roth and Kearns (2019) for an overview.

Within the literature, Kleinberg et al. (2017) examines extent to which different notions of fairness can or cannot be satisfied simultaneously by such algorithms, Dwork et al (2012), Kearns et al (2019) and Jung et al (2020) explores aspects of the trade-off between different notions of fairness and efficiency and Liang et al (2022) introduces a Pareto frontier that captures the optimal resolution of this trade-off for social planners with different preferences.

A common theme across all of these approaches is that the the set of covariates used to predict characteristics is assumed to be exogenously determined, with the goal being to examine how algorithms based on those covariates may or may not satisfy particular definitions of fairness. Here, the dataset itself is endogenous, with participation costs, data externalities and the incentives of the data collector all playing a role.

The source of unfairness in this paper, then, arises from the nature of the data collection process. There are a limited number of papers which analyse the contribution of the data collection process to inequality, of which the most relevant are Elzayn and Fish (2020), which models optimal data investment in both a monopolistic and competitive settings without externalities and Abebe (2019), which highlights the importance of data inequalities in the context of developing countries and proposes a search-engine driven approach to reduce those inequalities.

2 Model

In this section, I set out the nature of the information structure, payoffs of market participants and the timing of the game.

The information structure

Suppose there is a set, A, of agents, with an agent i is associated with a characteristic, $x_i \in \mathbb{R}$ and $|A| = n$. Let $\boldsymbol{x} = (x_1, ..., x_n)^T$ be a vector of characteristics which are joint normally distributed; $\boldsymbol{x} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \boldsymbol{\Sigma}(\boldsymbol{y}))$, where $\text{var}(x_i) = \xi_i^2$, $\text{cov}(x_i, x_j) = \xi_{ij}$ and $\boldsymbol{y}$ is the vector of secondary characteristics. The characteristic x_i is private information unless i sells their data to a service provider. Let $\boldsymbol{\Sigma}$ be an **information structure**.

Each agent has a secondary characteristic $y_i \in \{y_1, ..., y_l\}$, which correspond to a group, such that the group of agents $T_j \in \mathcal{T}$ denotes the set of all agents with secondary characteristic y_j. Assume that each agent's secondary characteristic is common knowledge. This secondary characteristic, which can be thought of as a demographic characteristic, like race or gender, determines the correlation structure. Specifically, assume that $\xi_{ij}(\boldsymbol{y}) = \xi_{st}(\boldsymbol{y})$ for all $i \in T_s$ and $j \in T_t$ and $\xi_i^2(\boldsymbol{y}) = \xi_s^2(\boldsymbol{y})$ for all $i \in T_s$.

A service provider ultimately wishes to estimate each x_i, either by buying i's data directly or using the data of other agents. Assume throughout that $\sum_j^{n-1} |\xi_{ij}| < \xi_i^2$

for all i and that all market participants are accurately informed about the parameter matrix $\boldsymbol{\beta}$ of the following linear regression:

$$x_i(\boldsymbol{\beta}, \varepsilon_i) = \mu_i + \sum_{j \neq i}^{n-1} \beta_{ij} x_j + \varepsilon_i \tag{1}$$

where ε_i is normally distributed such that $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma_i^2$, $\boldsymbol{\beta}$ is a zero-diagonal $n \times n$ matrix whose ijth entry is β_{ij} and $\mu_i \in \mathbb{R}$ measures some idiosyncratic (but known to the service provider) features of agent i. Note that the matrix $\boldsymbol{\beta}$ reflects the regression coefficients conditioning on every other variable, as opposed to the coefficient generated when one or more variable is omitted.

For ease of exposition, I assume throughout that $\xi_{ij}(\boldsymbol{y}) \geq 0$ for all i, j pairs, and thus, $\beta_{ij} \geq 0$, though the framework analysed here can easily encompass negative and positive correlations between characteristics.

Agents' data selling choice

The agents in the model choose whether to participate ($a_i = 1$) or not participate ($a_i = 0$) in the data selling game.

Let $p_i(\boldsymbol{a})$ be the price the agent receives from the service provider for their data and $\tau \geq 0$ capture the value the agents' receive from the service provider's ability to estimate their characteristic. Agent i's ex-post utility can be expressed as follows:

$$u_i(a_i = 1, \boldsymbol{a}_{-i}, p_i, c_i) = \begin{cases} p_i(\boldsymbol{a}) - c_i - \tau(x_i - d_i)^2 & \text{if } a_i = 1 \\ -\tau(x_i - d_i)^2 & \text{if } a_i = 0 \end{cases}.$$

The parameter c_i is the cost of participation (which captures, for example, the cost of time participating in data collection or linguistic and technical barriers to participation) which is assumed to be a function of i's secondary characteristic - specifically suppose each agent's cost is $c_i = c(y_s)$ for $i \in T_s$, where $c(y_s) > 0$.

The service provider

The service provider has two roles: they offer a subset of agents, $\boldsymbol{A_1} \subseteq \boldsymbol{A}$ (whose complement is $\boldsymbol{A_0}$), a take-it-or-leave it price, such that an agent i receives a price offer, p_i; and, based on the data they collect, they provide a targeted service based on their best estimate of i's characteristic.

Specifically, once the data is collected the service provider offers d_i to each agent, whether they participated in the data collection process or not. The expected payoff from the agent-specific action d_i is $\mathrm{E}[\sum_i (x_i - d_i)^2]$: the service provider wishes to provide a service which is a close as possible to the characteristic of the agent. Hence, the service provider's expected payoff can be summarised as follows:

$$\pi_S(\boldsymbol{a}) = \mathrm{E}[\sum_i (x_i - d_i)^2 | \boldsymbol{a}] - \sum_i p_i(\boldsymbol{a}).$$

Timing

The timing of the game is as follows:

1. The service provider chooses a vector of take-it-or-leave price offers, $\boldsymbol{p}$;
2. agents choose whether to participate in the data market or not, generating the vector $\boldsymbol{a}$;
3. the service provider chooses the service vector, $\boldsymbol{d}$.

3 Applications

There are a number of features of the model which capture aspects of particular data markets. Specifically, the data collection process is assumed to be costly, and so applications that best fit the model are those where data is actively elicited through some collection process, rather than being a byproduct of some other process, like data captured through tracking an agent's online behaviour. Furthermore, the service provider is assumed to be offering a service which is beneficial to the agent.

In this section, I analyse some applications that closely fit the model here. As is well-known, the use of data to make targeted decisions is becoming more and more prevalent and in many areas it is in its infancy. Hence, the sorts of policy questions this paper addresses will only become more relevant as time goes by, especially in areas such as public policy where personalised data techniques have yet to be adopted into the mainstream.

Healthcare

Personalised medicine, both at the level of groups and even individuals is becoming increasingly viable (see Armstrong, 2017 for analysis of the link between big data and personalised medicine and Vellekoop et al, 2022 for a literature review). The model is well-suited to such a situation because every agent benefits from a healthcare provider or drug manufacturer being able to accurately assess their response to a particular course of treatment. Furthermore, it is generally costly to participate in medical trials, and in many cases participants are paid non-trivial sums for that participation, with Fisher et al (2021) finding that the median payment for phase 1 clinical trial participation in the United States is $3,070.

Note also that healthcare is an area in which there are disparities in participation. For example, racial and ethnic minorities are less likely to participate in health research than non-minority individuals (see, for example, Milani et al, 2021) and the vast majority of trials take place in developed countries (Alemayehu, 2018). The model here seeks to understand how data markets contribute to these inequalities.

Product recommendations

Survey data and online reviews are used to supplement more passive data in improving recommendation systems online (see, for example, Sengupta, Srebro and Evans, 2018). In this setting, agents prefer to be given a recommendation which most accurately reflects their preferences, and those preferences may be elicited by participation in a (time) costly activity. Recommendation systems are becoming increasingly prominent in the online space, generating potentially large welfare gains (Zhang, 2017).

However, there is evidence that such systems are prone to a plethora of biases, see, for example, Chen et al (2020). Those authors identify a relevant bias for our purposes, which is the idea that products which are the most popular for the majority are more likely to be accurately estimated and, subsequently, be recommended to everyone. The model here seeks to capture the data markets which underpin such recommendation systems.

Employment markets

The use of data to match job seekers to employers is now common practice within that industry (Mezzanzanica and Mercorio, 2019). To the extent to which such job matching is preference-based, rather than skills-based, the model captures the notion that more information will lead to better job recommendations by platforms like LinkedIn or

Indeed, who would be the service provider in this setting.

Such job recommendation algorithms have shown biases - for example LinkedIn themselves have found that their recommendation algorithm was biased towards people who are more active on their platform, and such users were disproportionately represented in training data (Yin et al, 2021). The model here captures the effects of these inequalities, and illustrates how they can arise in a case where a job recommendation algorithm is trained on data acquired on an open data market.

4 Equilibrium

I characterise the equilibria of the data market and provide analysis of the pricing of data in those equilibria.

Definition

The strategy set $S = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{d}\}$ is an equilbrium if the following conditions hold:

1. $\boldsymbol{p} \in \text{argmax}_{\boldsymbol{p}\in\{0,1\}^n}\{\text{E}[\sum_i (x_i - d_i)^2|\boldsymbol{a}, \boldsymbol{p}] - \sum_i p_i(\boldsymbol{a})\}$;

2. $a_i \in \text{argmax}_{a\in\{0,1\}} u_i(a_i = a, \boldsymbol{a}_{-i}; \boldsymbol{p})\ \forall i$;

3. $d_i(\boldsymbol{a}) \in \text{argmax}_{d_i\in\mathbb{R}} \text{E}[(x_i - d_i)^2|\boldsymbol{a}]\ \forall i$.

I characterise the optimal pricing decision and the existence of equilibria below.

Data substitutability

Note that in any equilibrium, $d_i(\boldsymbol{a}) = \text{E}[x_i|\boldsymbol{a}]\ \forall i$. Let $\chi_i(\boldsymbol{a}) = \text{E}[\sum_i (x_i - \text{E}[x_i|\boldsymbol{a}])^2]$, $\chi(\boldsymbol{a}) = \sum_i \chi_i(\boldsymbol{a})$, $\gamma_i(\boldsymbol{a}_{-i}) = \chi(a_i = 0, \boldsymbol{a}_{-i}) - \chi(a_i = 1, \boldsymbol{a}_{-i})$ and $v_i(\boldsymbol{a})$ be the service provider's willingness to pay for i's data given the participation vector $\boldsymbol{a}$.

Lemma 1. *In any equilibrium,* $v_i(\boldsymbol{a}) = \gamma_i(\boldsymbol{a}_{-i})\ \forall i$.

The service provider is willing to pay for i's data to the extent to which it is informative of the characteristics of i and every other agent: that is the value of i's data in terms of the extent to which it enables the service provider to target its services more effectively.

Lemma 2. *Information acquisition regarding* i *is a substitute for information acquisition regarding* j*: that is,* $\gamma_i(a_j = 0, \boldsymbol{a}_{-i,j}) \geq \gamma_i(a_j = 1, \boldsymbol{a}_{-i,j})$.

If j sells their data, the total reduction in the mean-squared error associated with acquiring i's data decreases. This holds for two reasons: first, learning about i is directly informative of j, but also learning about i (weakly) decreases the service provider's uncertainty about every other agent, which also reduces the value of i's information to the service provider.

Existence of equilibrium

The preceding analysis suggests following Proposition relating to the existence of equilibrium:

Proposition 1. *The set of equilibria given an information structure* $\boldsymbol{\beta}$*,* $\Theta(\boldsymbol{\beta})$*, is always non-empty. Furthermore, the payoff to the service provider is the same in each equilibrium.*

Note that, in general, there are multiple equilibria. This follows because, as per Lemma 2, agents' participation choices affect the value of other agents' data, and as such the participation choice of every other agent, $\boldsymbol{a}_{-i}$, potentially affects the equilibrium action of i, a_i. Nevertheless, our results will hold for every possible equilibrium, unless otherwise stated.

The second statement in Proposition 1 follows from the fact that the service provider elicits participation with their choice of price vector. In which case, they choose potential equilibria which maximise their payoff, taking into account optimal agent participation decisions.

5 The market for information

I characterise the pricing in the data market utilising a network interpretation of the information structure. This interpretation provides insights into which agents participate in equilibrium and the price they receive for their data.

The value of data: a network interpretation

There are two effects of the service provider learning the characteristic of i: a direct effect whereby the service provider is able to precisely target i; and an indirect effect arising from the fact that knowing x_i also provides a more accurate estimate of the characteristics of other agents.

To measure both these effects, we need to consider the value of knowing the action profiles of different agents. Given equation (1) above, the following holds:

$$\boldsymbol{x} = M(\boldsymbol{\beta})[\mu + \boldsymbol{\varepsilon}] \tag{2}$$

where $M(\boldsymbol{\beta}) = [I - \beta]^{-1}$ and $\boldsymbol{\varepsilon}$ is the vector of realised error terms whose ith entry is ε_i. The matrix $M(\boldsymbol{\beta})$ is the Bonacich centrality matrix of the "network" implied by the regression coefficient matrix, $\boldsymbol{\beta}$, which under, this interpretation, is a weighted adjacency matrix.

It is well known (see Bonacich, 1987) that $M(\boldsymbol{a}) = \sum_{k=0}^{\infty} \boldsymbol{\beta}^k(\boldsymbol{a})$, which in turn implies that j's Bonacich centrality measures the sum of the (weighted) paths in G that begin at j, with the entries of matrix $M(a)$ then being the sum of paths that start at i and end at j. In this context, the centrality of an agent captures the fact that if j and k's characteristics are unknown learning i's characteristic is informative of j and k's characteristics directly, but it is also indirectly informative by virtue of the fact that j's characteristic is informative of k's (and vice versa).

We are interested in cases where a subset of agents have sold their data to an intermediary. Let $\hat{\boldsymbol{\beta}}(\boldsymbol{a})$ denote a regression coefficient matrix which only includes agents for whom the service provider has no information (that is, agents for whom $a_i = 0$). Further, let $\hat{\boldsymbol{x}}(\boldsymbol{a})$ be the vector of characteristics of those elements of the set A_0 and $\tilde{x}(\boldsymbol{a})$ be the vector of characteristics of those agents whose information the service provider has purchased. The following holds:

$$\hat{\boldsymbol{x}}(\boldsymbol{a}) = \hat{M}(a)[\hat{\mu}(\boldsymbol{a}) + \hat{\boldsymbol{\varepsilon}}(\boldsymbol{a}) + D(\boldsymbol{a})\tilde{x}(\boldsymbol{a})],$$

where $D(\boldsymbol{a})$ is the matrix of β_{ij}s between members of the set A_0 and every other agent and $\hat{\boldsymbol{\varepsilon}}(\boldsymbol{a})$ is a $|A_0| \times 1$ vector which includes the idiosyncratic error terms of those agents whose data has not been sold to the service provider.

As the service provider ultimately values the extent to which data reduces the mean-squared error of their estimate of the agents' characteristics, we will ultimately be interested in the matrix $S(\boldsymbol{a}) = \hat{M}^2(\boldsymbol{a})$. The ijth term in this matrix measures the sum of the even-numbered paths between i and j, and the matrix as a whole determines the value of the data of some agent i who is not currently participating in the data collection process, as shown by the following Lemma:

Lemma 3. *The following identity holds:*

$$S(a)\hat{\boldsymbol{\sigma}}^{\mathbf{2}}(\boldsymbol{a})\mathbf{1} = \sum_{i \in A_0}^{|A_0|} (x_i - \mathrm{E}[x_i|\boldsymbol{a}])^2. \tag{3}$$

Equation (3) forms the basis for our analysis in the remainder of this section.

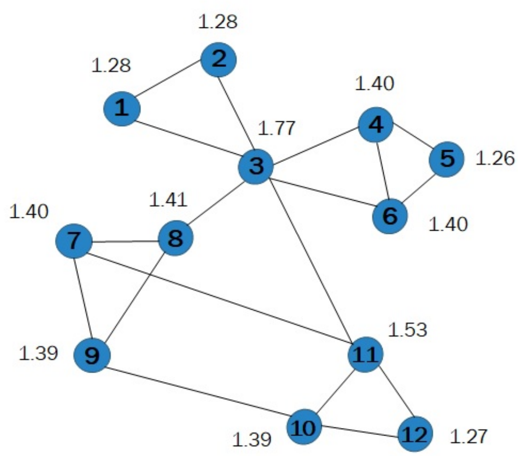

Figure 1: An information structure, where a link between two nodes represents $\beta_{ij} = 0.05$, and zero otherwise. The out of the circle numbers denote the even-path centrality of each agent within the information structure, assuming no agents choose to sell their data.

Equilibrium valuations and prices

We can use the above analysis to characterise the informativeness of learning an agent's characteristic. Define the $|A_0| \times 1$ $\phi(\boldsymbol{a}) := S(\boldsymbol{a})\hat{\boldsymbol{\sigma}}^{\mathbf{2}}(\boldsymbol{a})$, whose ith component is i's **even-path centrality** in the graph implied by $\boldsymbol{a}$ and i's **even-path intercentrality** as follows:

$$\theta_i(\boldsymbol{a}_{-i}) := \sum_{j\in A_0}^{|A_0|}[\phi_j(a_i = 0, \boldsymbol{a}_{-i}) - \phi_j(a_i = 1, \boldsymbol{a}_{-i})].$$

To return to the network interpretation, the even-path intercentrality of i, $\theta_i(\boldsymbol{a}_{-i})$, is a measure of the informativeness of i's data, which captures the reduction in the sum of paths of all lengths (weighted by the idiosyncratic variances of each agent) associated with learning x_i.[1] The following Theorem then holds:

Theorem 1. *Any equilibrium pair of participation and price vectors, $\boldsymbol{a}^*$ and $\boldsymbol{p}^*$, satisfy the following conditions:*

1) the service provider's valuation for agent i's data, $v_i(\boldsymbol{a}^*) = \theta_i(\boldsymbol{a}^*_{-i})$;

2) the price i receives for their data, $p^*_i(\boldsymbol{a}^*) = c_i - \tau\phi_i(a_i = 0, \boldsymbol{a}_{-i})$ if $a^*_i = 1$ and $p^*_i(\boldsymbol{a}^*) = 0$ otherwise;

3) $a^*_i = 1$ iff $\theta_i(\boldsymbol{a}^*_{-i}) \geq p^*_i(\boldsymbol{a}^*)$.

The informativeness measure, $\theta_i(\boldsymbol{a}^*_{-i})$, captures the extent to which learning x_i's characteristic will reduce the uncertainty associated with estimating the agents who the service provider does not have direct information on, weighed by the variance of the error term in the regression equation corresponding to those agents.

Theorem 1 establishes the principle that the value of an agent's data is not determined by their abstract informativeness (i.e. their informativeness when $\boldsymbol{a}_{-i} = \mathbf{0}$; $\theta_i(\mathbf{0})$), but moreover the extent to which they are informative of non-participating agents. Agents who are representative of high variance, non-participating agents are particularly valuable in equilibrium.

Furthermore, the Theorem provides a tractable account of how to measure repre-

[1] Put in these terms, there is a clear link between the service provider's problem and the key group problem within the network literature. In this literature, a central planner looks to remove a group of N agents from a network, with the aim of reducing aggregate centrality. See Bramoulle et al. (2016) for a recent survey.

sentativeness at the margin. For example, when $\sigma_i^2 = \sigma^2$ for all i, the measure $\theta_i(\boldsymbol{a}^*_{-i})$ is equal to $\frac{\phi_i(\boldsymbol{a}_{-i}, a_i=0)}{s_{ii}(\boldsymbol{a}_{-i}, a_i=0)} \sum_j s_{ji}$, which is an analogue of the concept of intercentrality found in Ballester et al. (2006) for the case where only even (weighted) path lengths of the graph implied by the participation vector $\boldsymbol{a}^*$ contribute to its calculated value. This compares with the price vector, which is determined by i's even-pathed centrality, $\phi_i(\boldsymbol{a}^*)$, in the same graph.

A benchmark

To aid in the understanding of Theorem 1, I consider two extreme cases; one where costs are sufficiently low such that every agent participates in the data market, and one where they are sufficiently large such that (at most) one agent participates:

Corollary 1. *Suppose $c_i = c$ for all i. The following two statements hold: (1) there exists a $\underline{c}$ such that if $c < \underline{c}$ then the is a unique equilibrium such that $p_i^*(\boldsymbol{a}^*) = c - \tau\sigma_i^2$ for all i; (2) there exists a $\bar{c}$ such that if $c > \bar{c}$, then, for an equilibrium with price vector $\boldsymbol{p}'$ and participation vector $\boldsymbol{a}'$, if $a_i' = 1$, then $p_i'(\boldsymbol{a}') = \theta_i(\boldsymbol{0})$.*

When costs are sufficiently small, every agent i participates and their data's value is determined by the variance of their idiosyncratic error term - in the case where the agent does not participate the service provider loses no information about any other agent, as they are all participating, and they are maximally informed about i's characteristic as the set-up will allow without actually having being sold i's data. Hence, each agent i receives the minimum possible price for their data, $c - \tau\sigma_i^2$.

When costs are sufficiently large, a single agent (at most) participates, and their data is valued according to its abstract informativeness. Furthermore, assuming $\sigma_i^2 = \sigma^2$ for all i, the agent who participates in equilbrium is the agent with the highest even-path intercentrality in the graph implied by the vector $\boldsymbol{a} = \mathbf{0}$.

Example

Suppose $\tau = 0.25$, $\beta_{ij} = 0.05$ or 0 (if there is no correlation between the two variables) for all i, j and $\sigma_i = 1$ for all i and the correlation structure is as depicted in Figure 2.

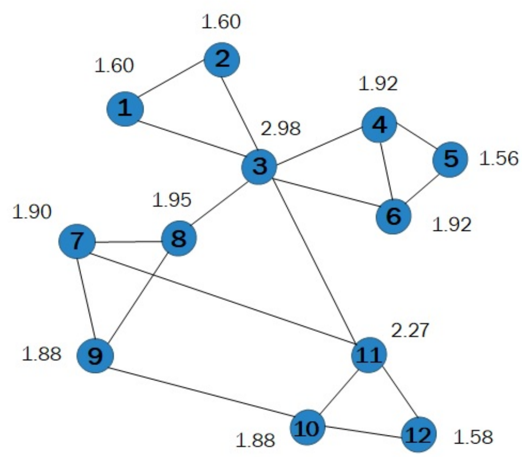

Figure 2: An information structure, with the out of circle values representing the even-path intercentralities of each node.

Suppose first that $c_i = 2$ for all i. Computing the value of the even-path intercentrality of each agent given the original information structure, it is clear that only a single agent can be induced to participate. In which case, it is optimal for the service provider to induce agent 3 to participate, as they have the highest intercentrality.

Now suppose that $c_3 = c_4 = c_8 = c_{11} = 0.5$ while $c_i = 3$ for all other i, which would imply that the four lower cost agents will participate in equilbrium. The above analysis makes it possible to characterise which agent receives the highest price for their data.

Calculating the even-path intercentralities for the case where the other participating agents participate, agent 11's is highest. This reflects the fact that while agent 3 is the most abstractly informative, they are less representative of non-participating agents than agent 11. Agent 11's data is therefore valued most highly by the service provider. Agent 11 also has the most even-path centrality by this metric - which implies that while their data is the most valuable, they receive the lowest price for their data of those agents who participate.

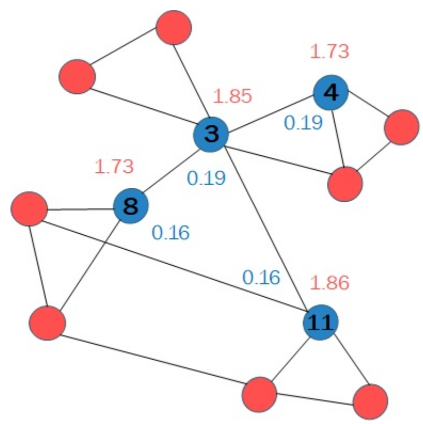

Figure 3: Prices (in blue text) and intercentralities (red) when 3, 4, 8 and 11 participate.

6 Equilibrium participation

In this section, I characterise the externalities present in data markets in which the end use of the data is beneficial to consumers, showing that the participation decision of one agent has the potential to crowd-out other agents. This free-riding effect has implications for how beneficial data leakages ultimately are for the service provider and data holders.

Crowding-out

Let a_i and a_j be strategic substitutes for a price vector $\boldsymbol{p}$ if:

$$\hat{u}_i(\boldsymbol{a}) = u_i(a_i = 1, a_j, \boldsymbol{a}_{-i,j}; \boldsymbol{p}) - u_i(a_i = 0, a_j, \boldsymbol{a}_{-i,j}; \boldsymbol{p})$$

is decreasing in a_j.

Proposition 2. *The participation decisions, a_i, a_j are strategic substitutes for all i, j pairs.*

Proposition 2 implies that when the service provider is offering a beneficial service, crowding-out occurs: as the service provider is more informed of an agent's characteristic when one of their peers chooses to participate, they have less incentive to participate. As such, agents who do not participate free-ride on those who do participate, receiving a higher payoff on account of the data leakages resulting from the sale of data by their peers.

Crowding-out and data leakages

The free-riding effect analysed above raises the possibility that environments where information leakage is high (i.e. ξ_{ij} is relatively high for all i, j) could result in the service provider receiving less information than in environments where information leakage is low. To illustrate this effect, I state the following definition:

Definition 1. An information structure $\boldsymbol{\Sigma}'$ has stronger data leakages than $\boldsymbol{\Sigma}$ if $\xi'_{ij} > \xi_{ij}$ and $\xi_{ii} = \xi'_{ii}$ for all i, j.

Given this definition, it is possible to consider the effect of crowding-out as data leakages become more pronounced. To do so, it also worth stating:

Definition 2. An information structure $\mathbf{\Sigma}'$ is (weakly) more informative in equilibrium than $\mathbf{\Sigma}$ if any pair of equilibrium participation vectors, $a^*(\mathbf{\Sigma}')$ and $a^*(\mathbf{\Sigma})$ are such that $\chi(a^*(\mathbf{\Sigma}')) > (\geq)\ \chi(a^*(\mathbf{\Sigma}))$.

Theorem 2. *Suppose $\mathbf{\Sigma}'$ has stronger data leakages than $\mathbf{\Sigma}$. Then, $\exists \bar{\tau} \in \mathbb{R}_+$ such that if $\tau > \bar{\tau}$ then $\mathbf{\Sigma}'$ is (weakly) more informative in equilibrium than $\mathbf{\Sigma}$, with the relation strict if, for all equilibrium participation vectors, there exists some i such that $a_i^*(\mathbf{\Sigma}') = 0$.*

Theorem 2 arises from the crowding-out effect. As per Theorem 1, agents only participate if their participation cost plus the benefit associated with a better targeted service is lower than the service provider's valuation for their data. As data leakages become more significant, the intrinsic value of selling data is reduced, as the service provider is more informed about non-participating agents. Therefore, fewer agents participate than they would do when data leakages are less significant.

As τ increases, this reduction in participation effect gets stronger, and it eventually dominates the fact that, for a fixed $\boldsymbol{a}$, the service provider is better informed when data leakages are stronger.

Hence, information structures which are more informative can result in the service provider being, somewhat paradoxically, less informed, because of the free-riding effect inherent to the case where the service provider offers a beneficial service.

7 Groups and data inequality

In order to analyse the datasets collected by the service provider, it is useful to put more structure on the link between costs and information structure. In doing so, it is possible to analyse which groups the service provider learns the most and least about, helping to provide an account of the informational inequalities that arise.

Preliminaries: the representative group game

Consider a special case of the model which will be referred to as a **representative group** game. Suppose that (as before) $\xi_{ij}(\boldsymbol{y}) = \xi_{st}(\boldsymbol{y})$ for all $i \in T_s$ and $j \in T_t$, but also that $\beta_{ss}(\boldsymbol{y}) = \beta_{tt}(\boldsymbol{y})$ for all s, t, $\beta_{tt}(\boldsymbol{y}) > \beta_{ts}(\boldsymbol{y})$ for $s \neq t$ and $\sigma_i^2 = \sigma_i$ for all i. Define $n_s = |T_s|$.

Finally, let $\hat{\chi}_j(\boldsymbol{a}) := \frac{\sum_{i \in T_j} \chi_i(\boldsymbol{a})}{n_j}$ be the average mean-square error of a group T_j for a participation vector $\boldsymbol{a}$. Assume throughout that for any equilibrium participation vector, $\boldsymbol{a}^*$,that $\hat{\chi}_i(\boldsymbol{a}^*) > 0$ for all i: there is at least one agent from each group who does not participate in equilibrium.

High cost groups

I start by analysing the extent to which the service provider is informed, on average, of the characteristics of a higher cost group:

Theorem 3. *Consider a representative group game where $\beta_{ks}(\boldsymbol{y}) \leq \beta_{kt}(\boldsymbol{y})$ for all $k \neq s, t$, $n_s = n_t$ and $c(y_s) > c(y_t)$. Then, for any equilibrium participation vector, $\boldsymbol{a}^*$, $\hat{\chi}_t(\boldsymbol{a}^*) > \hat{\chi}_s(\boldsymbol{a}^*)$.*

In the case where a group T_s is both weakly less informative and has a higher participation cost than group T_t then fewer of group T_s agents participate in equilibrium. As these are (weakly) less informative in the abstract, the service provider can always improve their payoff by inducing a higher proportion of individuals in low cost groups to participate than they do for a high cost group.

As such, members of high costs groups are disproportionately underrepresented in any collected dataset, which results in them receiving a lower quality service in equilibrium.

Minority groups

Now consider the case where there is a group, T_s, who are in the minority compared with some other group, T_t :

Theorem 4. *Consider a representative group game where* $\beta_{ks}(\boldsymbol{y}) \leq \beta_{kt}(\boldsymbol{y})$ *for all* $k \neq s, t$, *and* $c(y_s) = c(y_t)$. *If* $n_s < n_t$ *then* $\hat{\chi}_s(\boldsymbol{a}^*) < \hat{\chi}_t(\boldsymbol{a}^*)$.

Members of minority groups participate proportionately less in equilibrium and therefore the service provider is less informed about them on average than larger groups. This follows from the fact that, fixing the proportion of active agents in each group, a member of a minority group's data is less informative than the data from a member of a majority group, and so the service provider always prefers to negotiate with a greater proportion of the majority group.

This result is in line with results in the algorithmic fairness literature (see, e.g., Kleinberg et al, 2018) and the statistical discrimination literature more broadly (see Onuchic, 2022 for a survey) in which a utilitarian planner optimally chooses signals that are disproportionately informative about a majority group due to the increased value the planner receives from correctly estimating their type.

8 Efficiency, payoffs and inequality

This section analyses the efficiency of equilibria, as well as characterising the differences in agents' payoffs.

Efficiency

Consider the following definition of efficiency:

Definition 3. For a given participation vector $\boldsymbol{a}_{-i}$, it is efficient for i to participate (that is $a_i = 1$) iff:

$$\sum_{j=1}^{n} u_j(\boldsymbol{a}_{-i}, a_i = 1) + \pi_S(\boldsymbol{a}_{-i}, a_i = 1) \geq \sum_{j=1}^{n} u_j(\boldsymbol{a}_{-i}, a_i = 0) + \pi_S(\boldsymbol{a}_{-i}, a_i = 1)$$

Using this definition, I characterise the inefficiencies arising from the crowding-out effect analysed above.

Proposition 3. *An equilibrium $S = \{\boldsymbol{d^*}, \boldsymbol{a}^*, \boldsymbol{p^*}\}$ is inefficient if there exists an agent i such that $\theta_i(\boldsymbol{a}^*) + \tau\phi_i(\boldsymbol{a^*_{-i}}, a_i = 0) < c_i < (1+\tau)\theta_i(\boldsymbol{a}^*)$.*

When $c_i < (1+\tau)\theta_i(\boldsymbol{a}^*)$, the cumulative net payoff associated with $a_i = 1$ is positive and so it is efficient for i to participate. However, when $\theta_i(\boldsymbol{a}^*) + \tau\phi_i(\boldsymbol{a^*_{-i}}, a_i = 0) < c_i$, the service provider cannot induce i to participate while still making a profit, and as such S is inefficient.

Proposition 3 implies that there can be under-participation in equilibrium when the service provider offers a beneficial service. This follows from the fact that agents have an incentive to free-ride and do not take into account the positive externality that their participation decision has on non-participating agents. Note that $c_i > 0$ for some i is a necessary condition for there to be any inefficiency in this case: if $c_i = 0$, then every agent would participate.

Payoffs in equilibrium and inequality

To analyse the inequalities that arise in equilibrium, it is necessary to state the payoffs agents receive in equilibrium.

Theorem 5. *Agent i's utility in an equilibrium with participation vector* $\boldsymbol{a}^*$, $u_i(\boldsymbol{a}^*) = -\tau\phi_i(a_i = 0, \boldsymbol{a}^*_{-i})$. *Hence, if* $i, j \in T_s$, $a_i^* = 1$ *and* $a_j^* = 0$ *then* $u_j(\boldsymbol{a}^*, \boldsymbol{p}^*) \geq u_i(\boldsymbol{a}^*, \boldsymbol{p}^*)$ *with the inequality strict if* $\beta_{ij} \neq 0$.

Every agent, whether they participate or not, receives a payoff equal to their utility when they do not participate, which in turn is determined by the expected error the service provider makes in estimating their characteristic, which is given by the centrality measure, $\phi_i(\boldsymbol{a}_{-i}, a_i = 0)$. Hence, if, given an equilibrium in which the service provider learns less about i conditional on i not participating than they do about j, then i's payoff is lower than j's.

Furthermore, Theorem 5 states that a member of a group who does not participate is always better off than a member of that group who participates: if i participates in equilibrium then their utility reflects how informed the service provider would be about them if they were not to participate, which is necessarily less than how informed the service provider of the characteristic of a non-participating member of i's group. This difference in payoff reflects the positive externality which i's action exerts on non-participating agents.

Let $\bar{u}^j(\boldsymbol{a}^*) = \sum_{i \in T_j} \frac{u_i(\boldsymbol{a}^*)}{n_j}$. As discussed, there may be systematic differences in the extent to which the service provider is informed about different groups, which implies that there are differences in the utility they receive from data markets. For example, Theorem 3 directly implies the following result:

Corollary 2. *Consider a representative group game where* $\beta_{sk}(\boldsymbol{y}) = \beta_{tk}(\boldsymbol{y})$ *for all* k, $n_s = n_t$ *and* $c(y_s) > c(y_t)$. *Then, in any equilibrium,* $\bar{u}^s(\boldsymbol{a}^*) < \bar{u}^t(\boldsymbol{a}^*)$.

As shown above, the service provider is less informed about agents of group T_s than they are about group T_t agents when the former have a higher average cost than the latter. Therefore, Theorem 5 implies that agents in this group receive a lower payoff than group T_t agents do.

9 Privacy and data inequality

This section analyses the effect of a particularly effective privacy policy, where agents' data is decorrelated, such that the service provider only learns about them if they choose to their data.

A privacy policy

Consider the case where a social planner designs a decorrelation scheme, where the social planner commits to collecting the data on behalf of the service provider. The planner then sends out a distorted signal of the form $\hat{\boldsymbol{x}} = \Sigma^{-1}\boldsymbol{x}$.

This scheme allows for maximal privacy for those agents who do not participate in the data market - as such it is a maximally effective form of privacy legislation, allowing us to see the effect of weaker forms of such policies on utility and equality.

As Acemoglu et al (2022) point out, such a strategy can increase aggregate utility in data markets where the service is harmful to consumers, as it stops the provider learning the characteristic of any non-participating agents. In our setting, the implementation of such a scheme highlights the trade-off between equity and the benefits associated with data markets.

Privacy legislation and inequality

Let $u_i^d(a_i)$ denote i's utility under the privacy policy.

Proposition 4. *Suppose $\sigma_i = \sigma$ for all i. The outcome under the privacy policy is such that $\boldsymbol{u}_i^d(\boldsymbol{a_d}, \boldsymbol{p_d}) = \boldsymbol{u}_j^d(\boldsymbol{a_d}, \boldsymbol{p_d})$ for all i, j pairs.*

Proposition 4 highlights the fact that data leakages are the source of inequality in this model. Under the proposed privacy policy, one agent's participation decision does not provide any information on the characteristics of other agents. Hence, agents who participate receive the same utility as those who do not participate, and every agent receives exactly the same payoff.

Proposition 5. *For any outcome associated with any equilibrium without the privacy policy, $\boldsymbol{u}(a^*, p^*)$, $u_i(\boldsymbol{a^*}, \boldsymbol{p^*}) \geq u_i^d(\boldsymbol{a_d}, \boldsymbol{p_d})$ for all i, with the dominance strict for i if $\beta_{ij} \neq 0$ for at least one j for whom $a_j^* = 1$.*

For any information structure where there are some data leakages, the proposed data privacy policy reduces the utility of every agent. This follows because, as per Theorem 5, an agent's payoff in equilibrium is increasing in the service provider's ability to estimate their payoff. Under the privacy scheme, the service provider cannot learn anything about i's characteristic unless i participates. As a result, agents lose out under the policy, whether they sell their data or not.

The outcome with the proposed privacy policy is, at least weakly, worse for every group, and so it is Pareto dominated by the status quo outcome. At the same time, it is, in some sense, fairer than the status quo, in that every agent receives the same payoff.

Proposition 5 also shows that data leakages are the source of the inequalities highlighted in this setting: without data leakages, agents of each group receive the same payoff in equilibrium, though the service provider is still less informed about some groups than others.

Privacy policies in this setting, then, increase the fairness and equity associated with data markets but at the cost of the well-being of their participants.

10 Targeting inequality

Set-up

Suppose the social planner has a budget to buy data from agents directly. Let b_i denote the social planner's pricing offer to i and that they have a budget of $B > 0$, such that $\sum_i b_i \leq B$. Let $\hat{b}_s = \sum_{i \in T_s} b_i$ and assume that $B \geq c_i - \tau\phi(a_i = 0, \boldsymbol{a}^*)$ for some i where $a_i^* = 0$, so that the social planner can induce at least one extra agent to participate.

For simplicity, assume that the social planner buys data on a secondary data market in the following sense. The initial market takes place in the way described above, with the service provider collecting data and agents selling it without anticipating that the social planner will intervene.

The social planner then chooses a pricing vector, $\boldsymbol{b}$, taking as given the initial participation vector, $\boldsymbol{a}^*$. The social planner would then buy data on the secondary market, with some agents who did not participate in the initial market doing so in the secondary market. The social planner then hands over the data received on the secondary market to the service provider, who finally chooses $\boldsymbol{d}$, using the data obtained in both data markets.

The social planner's objective

Clearly, the optimal subsidy package depends upon the preferences of the social planner. For example, if the social planner were aiming to minimise the total error made by the service provider, the previous analysis tells us that they would choose $\boldsymbol{b}$ to minimise $\sum_i \phi_i(\boldsymbol{a}(b); \boldsymbol{a}^*, \boldsymbol{b})$.

A more interesting question arises when the social planner wishes to alleviate inequalities and/or increase the utility of particular groups. Again, the precise nature of the optimal intervention in this case will depend upon the weighting the social planner puts on equality versus aggregate utility. The analysis here will remain agnostic on the solution to this trade-off, instead showing how the framework can be used to identify how to increase the utility of a specific group disadvantaged by the initial data market.

Specifically, I analyse the optimal approach to increasing the utility of some group, T_s, i.e. the social planner solves the following minimisation problem

$$\min_{\boldsymbol{b}}\{\sum_{i\in T_s}\chi_i(\boldsymbol{a}(\boldsymbol{b}); \boldsymbol{a}^*)\}$$

subject to the constraint $\sum_i b_i \leq B$. I call T_s the social planner's "target group"

Examining the optimal solution to this problem provides insight into the more general trade-offs a social planner faces when trying to optimise their subsidy policy, taking into account the multiple inequalities which the data market generates.

Inequalities and targeting

The analysis above can be used to consider the benefit associated with an additional member of a group T_t being induced to participate in the data market on the aggregate utility of group T_s.

$$\hat{u}_t^s(\boldsymbol{a}) = \sum_{i \in T_t} [u_i(a_j = 1, \boldsymbol{a}_{-j}) - u_i(a_j = 0, \boldsymbol{a}_{-j})],$$

for some $j \in T_s$ and $a_j = 0$.

The matrix $\boldsymbol{S}(\boldsymbol{a})$ gives an insight into the level of substitution between i and j. For a participation vector $\boldsymbol{a}$, let:

$$\eta_{ij}(\boldsymbol{a}) = \phi_i(a_j = 0, \boldsymbol{a}_{\boldsymbol{-j}}) \frac{s_{ji}(a_j = 0, \boldsymbol{a}_{\boldsymbol{-j}})}{s_{jj}(a_j = 0, \boldsymbol{a}_{\boldsymbol{-j}})}.$$

I show in the Appendix that $\eta_{ij}(\boldsymbol{a})$ gives the reduction in the expected error the service provider makes when estimating i's characteristic associated with aquiring j's data for a participation vector, $\boldsymbol{a}_{\boldsymbol{-j}}$.

Proposition 6. *Consider a representative group game with participation vector* $\boldsymbol{a}$ *and suppose* $i \in T_i$, $j \in T_j$ *and* $k \in T_k$. *Then* $\hat{u}_i^j(\boldsymbol{a}) = \tau \eta_{ij}(\boldsymbol{a})$ *and so if* $\eta_{ij}(\boldsymbol{a}) > \eta_{ik}(\boldsymbol{a})$, *then* $\hat{u}_i^j(\boldsymbol{a}) > \hat{u}_i^k(\boldsymbol{a})$.

When $\eta_{ij}(\boldsymbol{a}) > \eta_{kj}(\boldsymbol{a})$, the reduction in the service provider's error in estimating i's characteristic associated with receiving j's data is greater than when they receive k's data when the participation vector is $\boldsymbol{a}$. Hence, buying data from a group j agent increases the average utility of group i agents more than buying data from a group k agent. This observation provides insight into the benefit of acquiring data from each group if the social planner's aim is to increase the utility of a particular group of agents.

Define:

$$\zeta_s(\boldsymbol{b}) := c_i - \tau\phi_i(\boldsymbol{a}(\boldsymbol{b}); \boldsymbol{a}^*),$$

which is the investment required in order to elicit an additional group T_s agent to participate when the investment vector is $\boldsymbol{b}$ and the equilibrium in the initial market is $\boldsymbol{a}^*$. The following statement holds:

Theorem 6. *Let T_i be the social planner's target group. If for all $\boldsymbol{b}$ such that $\boldsymbol{b}1^T = B$:*

$$\frac{\eta_{ij}(\boldsymbol{b})}{\eta_{ik}(\boldsymbol{0})} > \lfloor \frac{\zeta_k(\boldsymbol{0})}{\zeta_j(\boldsymbol{b})} \rfloor, \tag{4}$$

then for any $\boldsymbol{b}^*$, there exists a group, T_l such that $\bar{u}^l(\boldsymbol{b}^*) - \bar{u}^k(\boldsymbol{b}^*) > \bar{u}^l(0) - \bar{u}^k(0)$.

When equation (4) holds then, for any $\boldsymbol{b}^*$, the average utility of group j increases more than the average utility of group k due to the intervention of the social planner. Note that this result includes the case where $k = i$: that is, if (4) holds, then even when the social planner's only aim is to increase the utility of group T_i, the inequality between group j and group i increases after the intervention.

Equation (4) can hold when, for example, $c(y_i) > c(y_j)$: even when group i agents are more informative of their peers than group j agents, the fact the latter agents can be induced to participate at a cheaper price makes them a more valuable substitute overall.

Theorem 6 highlights the fact that achieving fairness and increasing welfare can be in conflict, even when the social planner is biased towards the interests of disadvantaged groups. In order to increase the payoff of such groups, it may be most effective to incentivise more advantaged groups to participate in the data market, but doing so has the effect of increasing inequality further.

11 Extensions

In this section, I explore some extensions to the model, providing an insight into potential further research.

Heterogenous service valuation

The preceding analysis was conducted on the basis that each agent had a common valuation for the service provided. It may be that groups differ on their valuation of that service. The most plausible account of this would be where some groups value privacy more than others. In order to make the payoffs between agents with different service valuations comparable, I re-define utility of an agent $i \in T_s$ as follows:

$$u_i(a_i = 1, \boldsymbol{a}_{-i}, p_i) = \begin{cases} p_i(\boldsymbol{a}) - c_i - (\tau_s + \tau)[(x_i - d_i)^2 - \xi_i^2] & \text{if } a_i = 1 \\ -(\tau_s + \tau)[(x_i - d_i)^2 - \xi_i^2] & \text{if } a_i = 0 \end{cases}.$$

where $\tau_s < 0$, and is assumed to be common knowledge. This formulation not only results in some agents gaining more net utility from the service than others, but opens up the possibility that some users receive negative utility from it. Let $\hat{\tau}_s = \tau_s + \tau$.

Proposition 7. *Suppose $\hat{\tau}_s < 0 < \hat{\tau}_t$, $j \in T_s$ and $i \in T_t$. Then, for any equilibrium S, $u_j(\boldsymbol{a}^*, p_j^*) < u_i(\boldsymbol{a}^*, p_i^*)$.*

When agents are sufficiently privacy conscious such that the service provider learning their type reduces their utility, they receive a lower payoff than a less privacy conscious group of agents. As such, the existence of privacy conscious agents makes privacy policies more appealing to a social planner: as privacy conscious groups are always less well off than their peers, the privacy policy is both better for aggregate utility and for equality concerns than it would be in the benchmark case analysed in the rest of the paper.

A more complex case to consider is one where an agent's desire for the service provider to learn their characteristic depends on the value of x_i. For example, suppose x_i is a measure of skill and the service provider assigns tasks based on that skill level, with more highly valued or paid tasks being optimally assigned to highly skilled individuals.

In this set-up, those with high values of x_i would benefit more from the service provider learning their value of x_i than those with lower values. Such a model would involve rich dynamics not captured here, as an agent's decision to not sell their data would be informative of their type in and of itself. As there are a number of real-life cases (e.g. benefits provision and some employment markets) which this model would capture, it would be fruitful to explore it in more detail in future work.

More general information structures

Throughout, I have assumed that agent characteristics are jointly normally distributed. Much of the analysis above regarding data inequality generalises to more general information structures, as long as these four properties hold:

1. **Monotonicity**: if $\boldsymbol{a} \geq \boldsymbol{a'}$, then $\chi_i(\boldsymbol{a}) \geq \chi_i(a')$ for all i.

2. **Group symmetry**: if two agents i and j are members of the same group, T_s, then $\chi_k(a_i = 1, a_j = 0, \boldsymbol{a_{-i,j}}) = \chi_k(a_i = 0, a_j = 1, \boldsymbol{a_{-i,j}})$ for all $\boldsymbol{a_{-i,j}}$ and k.

3. **Group informativeness**: if $i, j \in T_s$ and any $k \in T_t$ for $t \neq s$, then for all $\boldsymbol{a_{-j}}$:

$$\chi_i(a_i = 0, a_j = 0, \boldsymbol{a_{-j}}) - \chi_i(a_i = 0, a_j = 1, \boldsymbol{a_{-j}}) >$$
$$\chi_k(a_k = 0, a_j = 0, \boldsymbol{a_{-j}}) - \chi_k(a_k = 0, a_j = 1, \boldsymbol{a_{-j}}).$$

4. **Group submodularity**: if $i, j \in T_s$ and any $k \in T_t$ for $t \neq s$, then for all $\boldsymbol{a_{-j}}$:

$$\gamma_i(a_j = 0, \boldsymbol{a}_{-i,j}) - \gamma_i(a_j = 1, \boldsymbol{a}_{-i,j}) > \gamma_i(a_k = 0, \boldsymbol{a}_{-i,k}) - \gamma_i(a_k = 1, \boldsymbol{a}_{-i,k}).$$

Let ω_s^i denote the number of active members of group T_s excluding i and $\boldsymbol{a_{-s}}$ denote the activity vector excluding agents in group T_s.

Definition 4. Two groups, T_s and T_t, are **informationally identical** if for $i \in T_s$ and $j \in T_t$:

$$\chi(\omega_s^i = \omega, a_i = 0, \boldsymbol{a_{-s}}) - \chi(\omega_s^i = \omega, a_i = 1, \boldsymbol{a_{-s}}) =$$
$$\chi(\omega_t^j = \omega, a_j = 0, \boldsymbol{a_{-t}}) - \chi(\omega_t^j = \omega, a_j = 1, \boldsymbol{a_{-t}}).$$

for all $\boldsymbol{a_{-t}}$,$\boldsymbol{a_{-s}}$ and ω.

Proposition 8. *Suppose $c(y_s) > c(y_t)$, $n_s = n_t$ and that the groups T_s and T_t are informationally identical. Then, for any equilibrium participation vector, $\boldsymbol{a}^*$, $\hat{\chi}_t(\boldsymbol{a}^*) \geq \hat{\chi}_s(\boldsymbol{a}^*)$, with the inequality strict when $\boldsymbol{a}_i, \boldsymbol{a}_j \neq n_s$.*

As such, it is possible to generalise the notion of a group such that similar results regarding the level of information collected on such groups holds under more general information structures, and as such the policy implications highlighted here remain pertinent in more general cases.

Competition in the data market

In many real-life markets, the service provider would not collect data directly, but instead buy data from data intermediaries. Suppose that there are $m \geq 2$ data intermediaries, denoted by $\boldsymbol{D}$. A data intermediary i makes take-it-or-leave-it offers represented by the price vectors $\boldsymbol{p}^i$ and $\boldsymbol{q^i}$, whose jth component, q_j, represents the price charged to the service provider. I also suppose that data is non-rivalrous, and as such can be sold to multiple data intermediaries, incurring the same participation cost each time.

Let $\boldsymbol{P}$ denote a $n \times m$ matrix whose ijth entry is p^i_j, with Q defined analogously and $a^i_j = 1$ denote the case where an agent j sells their data to a data intermediary i. Define $\boldsymbol{\alpha}$ as a $n \times m$ matrix whose ijth entry is a^i_j. Finally, let $\boldsymbol{a_j}$ be the $m \times 1$ vector whose ith entry is a^i_j. A **competitive** equilibrium is defined as follows:

1. $\boldsymbol{q^i} \in \text{argmax}_{\boldsymbol{q^i} \in \{0,1\}^n} \{\sum_j q^i_j(\boldsymbol{Q^{-i}}; \boldsymbol{P}, \boldsymbol{\alpha})\} \ \forall i \in D$;

2. $\boldsymbol{p^i} \in \text{argmax}_{\boldsymbol{p^i} \in \{0,1\}^n} \{\sum_j q^i_j(\boldsymbol{Q^{-i}}; \boldsymbol{P}, \boldsymbol{\alpha}) - \sum_j p^i_j(\boldsymbol{P^{-i}}, \boldsymbol{\alpha})\} \ \forall i \in D$;

3. $a^i_j \in \text{argmax}_{\boldsymbol{a} \in \{0,1\}^m} u_j(\boldsymbol{a_j} = \boldsymbol{a}, \boldsymbol{\alpha}^{-j}; \boldsymbol{P}) \ \forall j$;

4. $d_j(\boldsymbol{a}) \in \text{argmax}_{d_j \in \mathbb{R}} \text{E}[(x_j - d_j)^2 | \boldsymbol{a}] \ \forall j$.

I state the following result regarding any competitive equilibrium:

Proposition 9. *In any competitive equilibrium with prices matrices $\boldsymbol{P^*}$and $\boldsymbol{Q^*}$ and participation matrix $\boldsymbol{\alpha}^*$, the following conditions are satisfied:*

1) the service provider pays intermediary i, $q^i_j(\boldsymbol{\alpha}^*) = \theta_j(\boldsymbol{\alpha}^*)$ for j's data;

2) if $\boldsymbol{a}_j \neq 0$, the price j receives for their data, $p_j(\boldsymbol{\alpha}^*) = c_j - \tau\phi_j(\boldsymbol{a_j} = \boldsymbol{0}, \boldsymbol{\alpha}^{-j})$.

3) agent j participates iff $\theta_j(\boldsymbol{\alpha}^{-j}) \geq p_i(\boldsymbol{\alpha}^*)$.

Proposition 9 highlights the point that competition in the data market does nothing to change the payoffs the agents receive for their data, and so does not change the fundamental differences in the allocation of payoffs present in our analysis.

The intuition for the result in Proposition 9 is consistent with Ichihashi (2021): as data is non-rivalrous, if two or more intermediaries set a price $p_j(\boldsymbol{\alpha}^*) \geq c_j - \tau\phi_j(a_j^* = 0, \boldsymbol{\alpha^{-j}})$ for j's data, i will sell to both firms. As j's data is homogenous, when more one data intermediary holds it, the resale value of that data is bid down to zero in equilibrium. Hence, only one data intermediary offers to buy any given agent's data in equilibrium, and so they receive the same price for their data as they do in the non-competitive equilibrium analyse in the main text.

12 Conclusion

As the data organisations hold about us becomes increasingly informative, the greater the extent to which those organisations can tailor their offerings and services to individuals and groups. This level of personalisation opens the door to a particular form of data inequality - between individuals who data collectors are well informed about, and individuals whose characteristics are relatively unknown.

The literature on such data inequalities have largely treated the dataset collected as a given, with the focus on how to deal with biases present within that data. Here, I analyse how incentives of participants in data markets determine those biases in the first place. To do so, this paper uses tools from the networks literature to characterise equilibrium prices, participation rates and the extent to which the data collector is informed about different market participants, highlighting the differences in the value of the data of different individuals.

The results here show how inequalities driven by data acquisition do not just relate to how data is used once it is collected: data markets themselves can contribute to data inequalities. Differences in the informativeness of a group's information, the size of the group and systematic differences in participation costs all contribute to these inequalities. Effectively alleviating these differences while maintaining the positive aspects of data markets of the type analysed here requires an understanding of the information structure.

As discussed in the extensions, there are a number of fruitful avenues for further research. Of the most promising, in my view, is the role data markets and the externalities they generate impact inequality in settings where the service provider's action is positive for agents with some characteristics and negative for others, with applications including criminal justice and hiring. Much of the literature on algorithmic fairness examines these cases, and the framework outlined here will generate new insights regarding these markets.

References

[1] Abebe, R. Designing algorithms for social good. Dissertation, 2019. https://ecommons.cornell.edu/handle/1813/70086.

[2] Alemayehu, C. M., Mitchell, G. K., and Nikles, J. Barriers for conducting clinical trials in developing countries- a systematic review. *International Journal for Equity in Health 17* (2018).

[3] Armstrong, S. Data, data everywhere: the challenges of personalised medicine. *BMJ 359* (2017).

[4] Babaioff, M., Kleinberg, R., and Leme, R. P. Optimal mechanisms for selling information, 2012.

[5] Ballester, C., Calvò-Armengol, A., and Zenou, Y. Who's who in networks. wanted: The key player. *Econometrica 74*, 5 (2006), 1403–1417.

[6] Bergemann, D., and Bonatti, A. Markets for information: An introduction. *Annual Review of Economics 11*, 1 (2019), 85–107.

[7] Bergemann, D., Bonatti, A., and Smolin, A. The design and price of information. *American Economic Review 108*, 1 (January 2018), 1–48.

[8] Bonacich, P. Power and centrality: A family of measures. *American Journal of Sociology 92*, 5 (1987), 1170–1182.

[9] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. Bias and debias in recommender system: A survey and future directions, 2020.

[10] Choi, J. P., Jeon, D.-S., and Kim, B.-C. Privacy and personal data collection with information externalities. *Journal of Public Economics 173* (2019), 113–124.

[11] Daron Acemoglu, Ali Makhdoumi, A. M. . A. O. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics (forthcoming)* (2022).

[12] de CorniÃšre, A., and de Nijs, R. Online advertising and privacy. *The RAND Journal of Economics 47*, 1 (2016), 48–72.

[13] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness, 2011.

[14] Eilat, R., Eliaz, K., and Mu, X. Optimal Privacy-Constrained Mechanisms. CEPR Discussion Papers 13536, C.E.P.R. Discussion Papers, Feb. 2019.

[15] Elzayn, H., and Fish, B. The effects of competition and regulation on error inequality in data-driven markets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 669â679.

[16] Engel, J. A. F. . C. Privacy as a public good. *65 Duke L.J. 385* (2015).

[17] Eso, P., and Szentes, B. Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies 74*, 3 (2007), 705–731.

[18] Gu, Y., Madio, L., and Reggiani, C. Exclusive Data, Price Manipulation and Market Leadership. Working Papers 202102, University of Liverpool, Department of Economics, Jan. 2021.

[19] Ichihashi, S. The economics of data externalities. *Journal of Economic Theory 196* (2021), 105316.

[20] Ichihashi, S. Dynamic Privacy Choices. Staff Working Papers 22-8, Bank of Canada, Mar. 2022.

[21] Jang, M., and Vorderstrasse, A. Socioeconomic status and racial or ethnic differences in participation: Web-based survey. *JMIR Res Protoc* (2019).

[22] Jung, C., Kannan, S., Lee, C., Pai, M. M., Roth, A., and Vohra, R. Fair prediction with endogenous behavior, 2020.

[23] Kearns, M., and Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* Oxford University Press, Inc., USA, 2019.

[24] Kearns, M., Roth, A., and Sharifi-Malvajerdi, S. Average individual fairness: Algorithms, generalization and experiments, 2019.

[25] Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic fairness. *AEA Papers and Proceedings 108* (May 2018), 22–27.

[26] Liang, A., Lu, J., and Mu, X. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation* (New York, NY, USA, 2022), EC '22, Association for Computing Machinery, p. 58â59.

[27] MacCarthy, M. Maccarthy, mark, new directions in privacy: Disclosure, unfairness and externalities (2011). 6 i/s: A journal of law and policy for the information society 425 (2011), available at ssrn: https://ssrn.com/abstract=3093301. *6 I/S: A Journal of Law and Policy for the Information Society 425* (2011).

[28] Mezzanzanica, M., and Mercorio, F. Big data for labour market intelligence: An introductory guide. Report, European Training Foundation, 2019.

[29] Montes, R., Sand-zantman, W., and Valletti, T. The value of personal information in online markets with endogenous privacy. *Management Science 65* (05 2018).

[30] Mulder, J., and de Bruijne, M. Willingness of online respondents to participate in alternative modes of data collection. *Survey Practice* (2019).

[31] National Academies of Sciences, E., and Medicine. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups.* The National Academies Press, Washington, DC, 2022.

[32] Onuchic, P. Recent contributions to theories of discrimination, 2022.

[33] Otufowora, A., Liu, Y., Young, H., Egan, K. L., Varma, D. S., Striley, C. W., and Cottler, L. B. Sex differences in willingness to participate in research based on study risk level among a community sample of african americans in north central florida. *Journal of immigrant and minority health 23*, 1 (February 2021), 19–25.

[34] Sadaf Arefi Milani, Michael Swain, A. O. L. B. C. C. W. S. Willingness to participate in health research among community-dwelling middle-aged and older adults: Does race/ethnicity matter? *Journal of Racial and Ethnic Health Disparities* (2021).

[35] Sengupta, N., Srebro, N., and Evans, J. Simple surveys: Response retrieval inspired by recommendation systems. *Social Science Computer Review 39*, 1 (2021), 105–129.

[36] Zhang, X.-C. Welfare properties of recommender systems, 2017.

Appendix

Proof of Lemmas 2 and 3

To prove Lemma 2, it is useful to utilise the result in Lemma 3. Hence, I prove the latter first. Recall that:

$$\hat{\boldsymbol{x}}(\boldsymbol{a}) = \hat{M}(a)[\hat{\mu}(\boldsymbol{a}) + \hat{\boldsymbol{\varepsilon}}(\boldsymbol{a}) + D(\boldsymbol{a})\tilde{x}(\boldsymbol{a})].$$

The vectors $D(\boldsymbol{a})\tilde{x}(\boldsymbol{a})$ and $\hat{\mu}(\boldsymbol{a})$ are fixed, and thus:

$$\mathrm{E}[\sum_i (x_i - \mathrm{E}[x_i|\boldsymbol{a}])^2] = [\hat{M}(a)\hat{\boldsymbol{\varepsilon}}(\boldsymbol{a})]^T[\hat{M}(a)\hat{\boldsymbol{\varepsilon}}(\boldsymbol{a})].$$

It then follows that:

$$[\hat{M}(a)\hat{\boldsymbol{\varepsilon}}(\boldsymbol{a})]^T[\hat{M}(a)\hat{\boldsymbol{\varepsilon}}(\boldsymbol{a})] = \hat{\boldsymbol{\varepsilon}}(\boldsymbol{a})^T\hat{M}^2(a)\hat{\boldsymbol{\varepsilon}}(\boldsymbol{a}),$$

and so the result holds in Lemma 3 holds. equity

Recall that $S(\boldsymbol{a}) = \sum_{k=0}^{\infty}\hat{\boldsymbol{\beta}}^{2k}(\boldsymbol{a})$ and $\hat{\sigma}_i^2(a_j = 0, \boldsymbol{a}_{-j}) = \hat{\sigma}_i^2(a_j = 1, \boldsymbol{a}_{-j})\ \forall i \neq j$. We know that:

$$\gamma_i(\boldsymbol{a}_{-i}) = \sum_{k\in A_0}^{|A_0|}[\phi_k(\boldsymbol{a}_{-i}, a_i = 0) - \phi_k(\boldsymbol{a}_{-i}, a_i = 1)].$$

The above expression for $\gamma_i(\boldsymbol{a}_{-i})$ implies that:

$$\gamma_i(\boldsymbol{a}_{-i}) = \sum_{k\in A_0}^{|A_0|}\sum_{j\in A_0}^{|A_0|}[s_{jk}(a_i = 0, a_{-i}) - s_{jk}(a_i = 1, a_{-i})\hat{\sigma}_k^2(a_i = 0, \boldsymbol{a}_{-i})].$$

To prove Lemma 2, I examine the change in the reduction in the sum of paths involved with the service provider learning i's characteristic when j is removed. Note that (as per Ballester et al, 2006), the following identity holds:

$$s_{jk}(a_i = 0, a_{-i}) - s_{jk}(a_i = 1, a_{-i}) = \frac{s_{ji}(a_i = 0, a_{-i})s_{ik}(a_i = 0, a_{-i})}{s_{ii}(a_i = 0, a_{-i})}, \tag{5}$$

and hence:

$$s_{ik}(a_i = 0, a_j = 0, a_{-i,j}) - s_{ik}(a_i = 1, a_j = 0, a_{-i,j}) \geq$$
$$s_{ik}(a_i = 0, a_j = 1, a_{-i,j}) - s_{ik}(a_i = 1, a_j = 1, a_{-i,j}).$$

This inequality directly implies the result in Lemma 2.

Proof of Proposition 1

For any given $\boldsymbol{a}$, $\exists d_i \in \text{argmax}_{d_i \in \mathbb{R}} \text{E}[(x_i - d_i)^2 | \boldsymbol{a}]\ \forall i$. To establish an equilibrium exists, it is then sufficient to show that the following participation and price vectors constitute an equilibrium:

$$\boldsymbol{a^*} = \text{argmax}_{a \in \{0,1\}^n} \sum_{i=1}^{n} [\chi_i(\boldsymbol{a}) + \tau \chi_i(a_i = 0, \boldsymbol{a}_{-i}) - c_i];$$

and $\boldsymbol{p^*(a^*)}$ whose ith component is $p_i^*(\boldsymbol{a}^*) = c_i - \tau \chi_i(a_i = 0, \boldsymbol{a}_{-i}^*)$ if $a_i^* = 1$ and 0 otherwise. First, note that no agent has a profitable deviation. If $a_i^* = 1$, then deviating and not participating gives a payoff equal to $-\tau \chi_i(a_i = 0, \boldsymbol{a}_{-i}^*) = u_i(a_i = 1, \boldsymbol{a_{-i}}, \boldsymbol{p^*})$. If $a_i^* = 0$, then deviating would yield a payoff of $-c_i$, which must be strictly less than $-\tau \chi_i(\boldsymbol{a}^*)$ by the definition of $\boldsymbol{a^*}$.

Now, we check that $\boldsymbol{p}^* \in \text{argmax}_{\boldsymbol{p} \in \{0,1\}^n} \{\text{E}[\sum_i (x_i - d_i)^2 | \boldsymbol{a}, \boldsymbol{p}] - \sum_i p_i(\boldsymbol{a})\}$. To see this, note that for some participation profile $\boldsymbol{a} \neq \boldsymbol{a^*}$ to be part of some equilibrium, $S = \{\boldsymbol{d}, \boldsymbol{a}, \boldsymbol{p}\}$, it must be the case that if $a_i = 1$, $p_i(\boldsymbol{a}) \geq c_i - \tau \chi_i(\boldsymbol{a}_{-i}, a_i = 0)$, otherwise i would deviate. Thus:

$$\pi_S(\boldsymbol{a}, \boldsymbol{p}) \leq \sum_{i=1}^{n} [\chi_i(\boldsymbol{a}) - c_i + \tau \chi_i(\boldsymbol{a}_{-i}, a_i = 0)] \leq \pi_S(\boldsymbol{a}^*, \boldsymbol{p}^*).$$

If some $S = \{\boldsymbol{d}, \boldsymbol{a}, \boldsymbol{p}\}$ constitutes an equilibrium, it must be the case that $\pi_S(\boldsymbol{a}, \boldsymbol{p}) = \pi_S(\boldsymbol{a}^*, \boldsymbol{p}^*)$, otherwise, there is a profitable deviation for the service provider, and therefore S would not be an equilibrium. Therefore, the payoff to the service provider is the same in each equilibrium.

Proof of Theorem 1

Note that $v_i(\boldsymbol{a^*}) = \gamma_i(\boldsymbol{a}_{-i})$. By equation (1):

$$\gamma_i(\boldsymbol{a}_{-i}) = \sum_{j \in A_0}^{|A_0|} [\phi_j(\boldsymbol{a}_{-i}, a_i = 0) - \phi_j(\boldsymbol{a}_{-i}, a_i = 1)],$$

and the right-hand side of this expression equals $\theta_i(\boldsymbol{a}^*_{-i})$ by definition.

Now, consider a candidate equilibrium price $p_i^*(\boldsymbol{a}^*) = c_i - \tau\chi_i(a_i = 0, \boldsymbol{a}^*_{-i})$. For any agent for whom $a_i = 1$ under some price vector, $\boldsymbol{p}$, and some participation vector, $\boldsymbol{a}$, it must be the case that:

$$p_i - c_i \geq -\tau\chi_i(a_i = 0, \boldsymbol{a}_{-i})$$

which implies that $p_i \geq -\tau\chi_i(a_i = 0, \boldsymbol{a}_{-i}) + c_i = p_i^*(\boldsymbol{a}^*)$. It follows that for any equilibrium participation vector, $\boldsymbol{a^*}$, in which $a_i^* = 1$, $p_i(\boldsymbol{a}^*) = p_i^*(\boldsymbol{a}^*)$: otherwise, the service provider can increase its profits by deviating to some lower price and still induce i to participate.

The identity in equation (1) establishes $\chi_i(a_i = 0, \boldsymbol{a}_{-i}) = \phi_i(a_i = 0, \boldsymbol{a}_{-i})$, and so the identity in the second statement in the Theorem holds.

The third statement follows from the fact that if $\theta_i(\boldsymbol{a}^*_{-i}) < p_i^*(\boldsymbol{a}^*)$, then $v_i(\boldsymbol{a^*}) = \theta_i(\boldsymbol{a}^*_{-i}) < p_i^*(\boldsymbol{a}^*)$, and so it is optimal for the service provider to deviate and set $p_i(\boldsymbol{a}^*) = 0 < p_i^*(\boldsymbol{a}^*)$, with the final inequality holding because $\theta_i(\boldsymbol{a}^*_{-i}) > 0$.

Proof of Corollary 1

By Theorem 1, $p_i^*(\boldsymbol{a}^*) = c_i - \tau\chi_i(a_i = 0, \boldsymbol{a}^*_{-i})$ in any equilibrium. Suppose that $c < \tau\sigma_i^2$ $\forall i$ then in the only equilibrium it must be that every agent participates, and receives a price $p_i^*(\boldsymbol{a}^*) = c - \tau\sigma_i^2$. To see that such a price vector uniquely induces $\boldsymbol{a^*} = \mathbf{1}$ when $c < \tau\sigma_i^2$ for all i, consider another participation vector $\boldsymbol{a} \neq \mathbf{1}$. Then, as $c < \tau\sigma_i^2$ and $\sigma_i^2 \leq \chi_i(a_i = 0, \boldsymbol{a}_{-i})$, any agent for whom $a_i = 0$ would deviate and receive a higher payoff.

Furthermore, $p_i^*(\boldsymbol{a}^*) = c - \tau\sigma_i^2$ is the lowest possible price to ensure every agent participates. Hence the proposed equilibrium in which $\boldsymbol{a^*} = \mathbf{1}$ and $p_i^*(\boldsymbol{a}^*) = c - \tau\sigma_i^2$ for all i is indeed a unique equilibrium where $v_i^*(\boldsymbol{a}^*) = \sigma_i^2$. This equilibrium pertains when $c < \tau\sigma_i^2$ for all i, and hence the first statement in the Corollary holds.

Now, suppose that $\gamma_i(\mathbf{0}) - c - \tau\chi_i(a_i = 0, \mathbf{0}) \leq 0$ for all i. For an agent i to be induced to participate $p_i \geq c - \tau\chi_i(a_i = 0, \mathbf{0})$. Furthermore, $\chi_j(a_i = 0, \mathbf{0}) \geq \chi_j(a_i = 1, \mathbf{0})$ with the inequality strict when $\theta_j(a_i = 0, \mathbf{0}) > \theta_j(a_i = 1, \mathbf{0})$. It follows that, if $a_i^{'} = 1$, then $p_i^{'}(\boldsymbol{a'}) = \theta_i(\mathbf{0})$ otherwise, i would deviate, preferring not to participate.

Recall that these conditions hold when $\gamma_i(\mathbf{0}) - c - \tau\chi_i(a_i = 0, \mathbf{0}) \leq 0$ for all i, and so exists a $\bar{c}$ such that if $c > \bar{c}$ such that any equilibrium is as described.

Proof of Proposition 2

Note that $u_i(a_i = 1, a_j, \boldsymbol{a}_{-i,j}; \boldsymbol{p}) = p_i(\boldsymbol{a}) - c_i$ for $a_j = 0$ and:

$$u(a_i = 0, a_j, \boldsymbol{a}_{-i,j}; \boldsymbol{p}) = -\tau\chi_i(a_j, \boldsymbol{a}_{-i,j})$$

As shown in the proof of Lemma 2, $\chi_i(a_j = 1, \boldsymbol{a}_{-i,j}) \leq \chi_i(a_j = 0, \boldsymbol{a}_{-i,j})$. As i's utility when $a_i = 1$ is independent of $\chi_i(a_j, \boldsymbol{a}_{-i,j})$, it follows $u(a_i = 1, a_j, \boldsymbol{a}_{-i,j}) - u(a_i = 0, a_j, \boldsymbol{a}_{-i,j})$ is weakly decreasing in a_j, and is strictly decreasing when $\chi_i(a_j = 1, \boldsymbol{a}_{-i,j}) < \chi_i(a_j = 0, \boldsymbol{a}_{-i,j})$.

Proof of Theorem 2

First, consider the information structure $\boldsymbol{\Sigma}$. Suppose that $\bar{\tau}$ is such that there is a unique equilibrium where $\boldsymbol{a^* = 1}$ and $p_i^*(\boldsymbol{a}^*) = c_i - \bar{\tau}\sigma_i^2$ for all i, and furthermore that $c_i - \bar{\tau}\sigma_i^2 = \sigma_i^2$ for at least one i. This final equality guarantees that the service provider is indifferent between buying and not buying i's data, and so if σ_i^2 were to increase, then they deviate and set $p_i^*(\boldsymbol{a}^*) = 0$ and induce i not to participate. In this case, $\chi(\boldsymbol{a^*}(\Sigma), \bar{\tau}) = 0$.

I partition the covariance matrix Σ such that:

$$\Sigma = \begin{array}{c|c} \xi_{ii} & \Sigma_{ij} \\ \hline \Sigma_{ji} & \Sigma_{jj} \end{array},$$

where $\boldsymbol{x}_j$ is the $(n-1)\times 1$ of the characteristics of non-i agents. Note that, by standard results, $x_i|\boldsymbol{x_j}$ is distributed normally with mean $\mu_i + \Sigma_{ij}\Sigma_{jj}^{-1}(\boldsymbol{x_j} - \boldsymbol{\mu_j})$ and variance $\xi_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}$. If $\xi_{ij}^{'} > \xi_{ij}$ and $\xi_{ii}^{'} = \xi_{ii}$ for all i, j, this then implies that $\sigma_j^{'} < \sigma_j$ for all j.

This final inequality implies that $\frac{c_i}{1+\bar{\tau}} > (\sigma_i^{'})^2$, which implies that the service provider is unwilling to induce $a_i = 1$ under the information structure $\Sigma^{'}$. It follows that for $\tau = \bar{\tau}$, the inequality in the Theorem holds strictly, with $\chi(\boldsymbol{a}(\Sigma^{'}), \bar{\tau}) > 0$ for any $\boldsymbol{a}(\Sigma^{'})$ which is an equilibrium under the information structure $\Sigma^{'}$.

As τ increases, the inequality $c_i - \tau\sigma_i^2 \leq \sigma_i^2$ continues to hold, and so $\chi(\boldsymbol{a^*}(\Sigma), \tau) = 0$ for all $\tau > \bar{\tau}$. Hence, $\chi(a^*(\boldsymbol{\Sigma}), \tau)(\geq)\ \chi(a^*(\boldsymbol{\Sigma}^{'}), \tau)$ for all $\tau > \bar{\tau}$ and any pair of equilibrium participation vectors, with the inequality strict if $a^*(\boldsymbol{\Sigma}^{'}) \neq \boldsymbol{1}$.

Proof of Theorem 3

Suppose that, for some equilibrium with participation vector $\boldsymbol{a}^*$ and price vector $\boldsymbol{p}^*$, $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$. Given that $\beta_{ks}(\boldsymbol{y}) \leq \beta_{kt}(\boldsymbol{y})$ for all $k \neq s, t$, $n_s = n_t$ and $\sigma_i = \sigma$ for all i, this implies that $\sum_{i \in T_s} a_i^* \geq \sum_{i \in T_t} a_i^*$ - that is, a greater proportion of type ts are active in the proposed equilibrium than type s agents. Note also that, by the assumption, $\boldsymbol{a}^*$, $a_i^* = 0$ for some $i \in T_k$ and all k.

The following Lemma holds:

Lemma 4. *When* $\sum_{i \in T_s} a_i^* \geq \sum_{i \in T_t} a_i^*$, $\beta_{sk}(\boldsymbol{y}) \leq \beta_{tk}(\boldsymbol{y})$ *for all* $k \neq s, t$, $n_s = n_t$ *and* $c(y_s) > c(y_t)$, *the intercentrality of some agent,* $i \in T_t$, $\theta_i(\boldsymbol{a}^*) \geq \theta_j(\boldsymbol{a}^*)$, *where* $j \in T_s$.

Proof. By definition:

$$\theta_i(\boldsymbol{a}_{-i}) := \sum_{j \in A_0}^{|A_0|} [\phi_j(a_i = 0, \boldsymbol{a}_{-i}) - \phi_j(a_i = 1, \boldsymbol{a}_{-i})]$$

we know that $\phi_j(a_i = 0, \boldsymbol{a}_{-i})$ is the jth entry of the matrix $\boldsymbol{S}(a)\boldsymbol{\sigma}^2$, and hence $\phi_j(a_i = 0, \boldsymbol{a}_{-i}) = \sum_j s_{ji}(a_i = 0, \boldsymbol{a}_{-i})\sigma^2$. As $\boldsymbol{S}(a) = \sum_{k=1}^{\infty} \boldsymbol{\beta}^k(a)$, then:

$$\theta_i(\boldsymbol{a}_{-i}) = \phi_i(a_i = 0, \boldsymbol{a}_{-i}) + \sum_{z=0}^{\infty} (\sum_{j \neq i} \sum_{k \neq i} g_{j(i)k}^{[z]}), \tag{6}$$

where $g_{j(i)k}^{[z]}$ is the value of the (weighted) paths of length z that begin at j and end at k. As $\beta_{ks}(\boldsymbol{y}) \leq \beta_{kt}(\boldsymbol{y})$ for all $k \neq s, t$ and $\beta_{ss} = \beta_{tt}$, it follows that when $\sum_{i \in T_s} a_i^* = \sum_{i \in T_t} a_i^*$, it must be that for $i \in T_t$ and $j \in T_s$, $\sum_{z=0}^{\infty} g_{r(i)k}^{[z]} \geq \sum_{z=0}^{\infty} g_{r(j)k}^{[z]}$ for all $r, k \notin T_s, T_t$;

$$\sum_{z=0}^{\infty} [g_{r(i)k}^{[z]} - g_{r(j)k}^{[z]}] \geq \sum_{z=0}^{\infty} [g_{r(j)k'}^{[z]} - g_{r(i)k'}^{[z]}]$$

for all $k \in T_t$ and $k^{'} \in T_s$ and:

$$\sum_{z=0}^{\infty}[g^{[z]}_{r(i)k} - g^{[z]}_{r(j)k}] \geq \sum_{z=0}^{\infty}[g^{[z]}_{r'(j)k} - g^{[z]}_{r'(i)k}]$$

for all $r \in T_t$ and $r^{'} \in T_s$, with each of these inequalities holding with equality iff $\beta_{sk}(\boldsymbol{y}) = \beta_{tk}(\boldsymbol{y})$ for all $k \neq s, t$. When $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$ then $\phi_i(a_i = 0, \boldsymbol{a}_{-i}) \geq \phi_j(a_j = 0, \boldsymbol{a}_{-j})$ for $i \in T_t$ and $j \in T_s$.

Notice that precisely the same weak inequalities all hold strictly when $\sum_{i \in T_s} a_i^* > \sum_{i \in T_t} a_i^*$. Hence, the Lemma holds. $\square$

Recall that $p_i^* = c_i - \tau\chi_i(\boldsymbol{a}^*)$ in equilibrium. It then follows that the service provider has an incentive to deviate, such that they set $p_j = 0$ for some $j \in T_s$ for whom $a_j^* = 1$ and set $p_i = c_i - \tau\chi_i(\boldsymbol{a}^*)$. Noting that $\chi_i(\boldsymbol{a}^*) < \chi_j(\boldsymbol{a}^*)$ as $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$ and $c(y_s) > c(y_t)$ it follows that $p_i < p_j^*$, and so the proposed deviation is profitable, yielding a contradiction.

Proof of Theorem 4

Suppose that, for some equilibrium with participation vector $\boldsymbol{a^*}$ and price vector $\boldsymbol{p^*}$, $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$. As $n_s < n_t$, $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$ implies that $\sum_{i \in T_s} \frac{a_i^*}{n_s} > \sum_{i \in T_t} \frac{a_i^*}{n_t}$.

Lemma 5. *When* $\sum_{i \in T_s} \frac{a_i^*}{n_s} \geq \sum_{i \in T_t} \frac{a_i^*}{n_t}$, $\beta_{sk}(\boldsymbol{y}) \leq \beta_{tk}(\boldsymbol{y})$ *for all* $k \neq s, t$, $n_s < n_t$ *and* $c(y_s) = c(y_t)$, *the intercentrality of some agent,* $i \in T_t$, $\theta_i(\boldsymbol{a^*}) > \theta_j(\boldsymbol{a^*})$, *where* $j \in T_s$.

Proof. A formula for the intercentrality of an agent i is given in equation (6). Suppose that $\sum_{i \in T_s} \frac{a_i^*}{n_s} = \sum_{i \in T_t} \frac{a_i^*}{n_t}$ (which, as per the above, is not true in this case) and $\beta_{sk}(\boldsymbol{y}) = \beta_{tk}(\boldsymbol{y})$ for all $k \neq s, t$. Clearly, as $n_s < n_t$ and $\beta_{tt}(\boldsymbol{y}) > \beta_{ts}(\boldsymbol{y})$ for $s \neq t$, the following inequalities hold for $i \in T_t$ and $j \in T_s$:

$$\sum_{z=0}^{\infty} g_{r(i)k}^{[z]} - \sum_{z=0}^{\infty} g_{r(j)k}^{[z]} > 0$$

for all $r, k \notin T_s, T_t$;

$$\sum_{z=0}^{\infty} [g_{r(i)k}^{[z]} - g_{r(j)k}^{[z]}] - \sum_{z=0}^{\infty} [g_{r(j)k'}^{[z]} - g_{r(i)k'}^{[z]}] > 0$$

for all $k \in T_t$ and $k^{'} \in T_s$ and:

$$\sum_{z=0}^{\infty} [g_{r(i)k}^{[z]} - g_{r(j)k}^{[z]}] - \sum_{z=0}^{\infty} [g_{r'(j)k}^{[z]} - g_{r'(i)k}^{[z]}] > 0$$

for all $r \in T_t$ and $r^{'} \in T_s$. These inequalities continue to hold if $\sum_{i \in T_s} \frac{a_i^*}{n_s} > \sum_{i \in T_t} \frac{a_i^*}{n_t}$ and/or $\beta_{ks}(\boldsymbol{y}) \leq \beta_{kt}(\boldsymbol{y})$ for all $k \neq s, t$: the left-hand side of each of the above inequalities are increasing in $\sum_{i \in T_s} \frac{a_i^*}{n_s} - \sum_{i \in T_t} \frac{a_i^*}{n_t}$ and $\beta_{kt}(\boldsymbol{y}) - \beta_{ks}(\boldsymbol{y})$ for all $k \neq s, t$.

When $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$, it must be that when $n_s < n_t$, $\phi_i(a_i = 0, \boldsymbol{a}_{-i}) \geq \phi_j(a_j = 0, \boldsymbol{a}_{-j})$ for $i \in T_t$ and $j \in T_s$, and hence by equation (6), $\theta_i(\boldsymbol{a}^*) > \theta_j(\boldsymbol{a}^*)$. □

Recall that $c_i = c_j$ and $\phi_i(a_i = 0, \boldsymbol{a}_{-i}) \geq \phi_j(a_j = 0, \boldsymbol{a}_{-j})$ for $i \in T_t$ and $j \in T_s$, and therefore $p_i^* = c_i - \tau\chi_i(\boldsymbol{a}^*)$.

It then follows that the service provider has an incentive to deviate from the proposed equilibrium with participation vector $\boldsymbol{a}^*$, such that they set $p_j = 0$ for some $j \in T_s$ for whom $a_j^* = 1$ and set $p_i = c_i - \tau\chi_i(\boldsymbol{a}^*)$. Noting that $\chi_i(\boldsymbol{a}^*) < \chi_j(\boldsymbol{a}^*)$ as $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$ and $c(y_s) > c(y_t)$ it follows that $p_i < p_j^*$, and so the proposed deviation is profitable, yielding a contradiction.

Proof of Proposition 3

Suppose $a_i^* = 0$ for some equilibrium vector $\boldsymbol{a^*}$. The total gross increase in payoff associated with the payoff vector $\boldsymbol{a'}$ where $a_i^{'} = 1$ and $a_j^{'} = a_j^*$ where $j \neq i$ is equal to $(1+\tau)\theta_i(\boldsymbol{a}^*)$ by Theorem 1. Hence, if $c_i < (1+\tau)\theta_i(\boldsymbol{a}^*)$,then:

$$\sum_{j=1}^{n} u_j(\boldsymbol{a'}) + \pi_S(\boldsymbol{a'}) \geq \sum_{j=1}^{n} u_j(\boldsymbol{a^*}) + \pi_S(\boldsymbol{a^*}).$$

However, as shown by Theorem 1, the maximum price the service provider is willing to pay i in equilibrium is $\theta_i(\boldsymbol{a}^*)$. Hence, if $c_i - \tau\phi_i(a_i = 0, \boldsymbol{a^*_{-i}}) > \theta_i(\boldsymbol{a}^*)$, then i cannot be induced to participate in equilibrium. Rearranging, we get the first inequality in the Proposition, and hence S is inefficient.

Proof of Theorem 5

The first statement in Theorem 5 follows from Theorem 1: as $p_i^*(\boldsymbol{a}^*) = c_i - \tau\phi_i(\boldsymbol{a}_{-i})$ if $a_i^* = 1$, then $u_i(a_i = 0, \boldsymbol{a^*}) = -\tau\phi_i(a_i = 0, \boldsymbol{a^*_{-i}})$ if $a_i^* = 1$. Furthermore, if $a_i^* = 0$ then this equality holds immediately.

For the second statement, note that $\chi_j(a_i = 0, \boldsymbol{a^*_{-i}}) \geq \chi_j(a_i = 1, \boldsymbol{a^*_{-i}})$, with the inequality strict if $\beta_{ij} > 0$. By definition, if $i, j \in T_s$ and $a_j^* = 0$ then $\chi_i(a_i = 0, \boldsymbol{a^*_{-i}}) = \chi_j(a_i = 0, \boldsymbol{a^*_{-i}})$. Hence:

$$u_i(\boldsymbol{a^*}) = -\tau\chi_i(\boldsymbol{a^*}) \leq -\tau\chi_j(\boldsymbol{a^*}) = u_j(\boldsymbol{a^*})$$

with the inequality strict if $\beta_{ij} > 0$.

Proof of Corollary 2

First note that, by Theorem 1, $\beta_{sk}(\boldsymbol{y}) = \beta_{tk}(\boldsymbol{y})$ for all k, $n_s = n_t$, $c(y_s) > c(y_t)$ implies that $\sum_{i \in T_t} a_i^* > \sum_{i \in T_s} a_i^*$, and specifically, $\sum_{i \in T_t} a_i^* - \sum_{i \in T_s} a_i^* \geq 1$. Suppose $a_i^* = 1$ and $i \in T_t$. It then follows that $\chi_i(a_i = 0, \boldsymbol{a^*_{-i}}) \geq \chi_j(a_j^* = 0, \boldsymbol{a^*_{-j}})$ for any $j \in T_s$ with $a_j^* = 0$. Given the condition on equilibrium prices in Theorem 1, it follows that $u_i(\boldsymbol{a^*}) \geq u_j(\boldsymbol{a^*}) > u_k(\boldsymbol{a^*})$ for some $k \in T_s$ where $a_k^* = 0$. The last inequality follows from Theorem 5.

Now consider some $l \in T_t$ and $a_i^* = 0$. By Theorem 5, $u_l(\boldsymbol{a^*}) > u_i(\boldsymbol{a^*})$. Hence, for any pair of agents (i, j), with $i \in T_t$ and $j \in T_s$ pairs, it follows that $u_i(\boldsymbol{a^*}) \geq u_j(\boldsymbol{a^*})$, with the inequality strict for some pairs. The Corollary follows immediately from this observation.

Proof of Proposition 4

First, I state the following Lemma:

Lemma 6. *Under the privacy policy, for all* $\boldsymbol{a}$*:*

$$\chi_i^d(\boldsymbol{a}) = \begin{cases} \chi_i(\boldsymbol{0}) & \text{if } a_i = 0 \\ 0 & \text{if } a_i = 1 \end{cases}$$

Proof. First note that $\mathrm{E}[\hat{\boldsymbol{x}}\boldsymbol{x}^T] = \mathrm{E}[\boldsymbol{x}\boldsymbol{x^T}]\Sigma^{-1} = I$. Thus, there is no correlation between x_i and x_j for all $j \neq i$. When $a_i = 1$, the service provider learns x_i with complete accuracy, and thus $\chi(\boldsymbol{a}) = 0$. □

The service provider's error when estimating x_i, $\phi_i(a_i = 0, \boldsymbol{a_{-i}}) \geq \phi_i(a_i = 0, \boldsymbol{0})$ with the inequality strict if $\beta_{ij} \neq 0$ for some $a_j^* = 1$. Lemma 5 implies that $\phi_i^d(a_i = 0, \boldsymbol{a_{-i}}) =$

$\phi_i^d(a_i = 0, \mathbf{0})$

By Theorem 5, $u_i(\boldsymbol{a^*}) = -\tau\phi_i(a_i = 0, \boldsymbol{a_{-i}})$. The preceding analysis implies that for any equilibrium participation vector $\boldsymbol{a^*}$, $-\tau\phi_i(a_i = 0, \boldsymbol{a^*_{-i}}) \geq -\tau\phi_i^d(a_i = 0, \boldsymbol{a})$ for any possible $\boldsymbol{a}$, including an equilibrium participation vector under the privacy policy, $\boldsymbol{a_d}$.

Proof of Proposition 5

Note that by Lemma 5 and Theorem 5, $u_i^d(\boldsymbol{a}_d) = -\tau\chi_i(\mathbf{0})$ for all i any equilibrium under the privacy policy. Hence, if $\sigma_i = \sigma$ for all i, $u_i^d(\boldsymbol{a}_d) = u$ for all i.

When $\boldsymbol{\beta} = 0$, if $a_i = 0$ then $\chi_i(a_i = 0, \boldsymbol{a_{-i}}) = \chi_i(\mathbf{0})$ for all a_{-i}. In this case, $u_i(\boldsymbol{a}^*, \boldsymbol{p^*}) = u$ for any equilibrium. If $\beta_{ij} > 0$ for some $j \in A_1$, then $u_i(\boldsymbol{a}^*, \boldsymbol{p^*}) > u$ for any equilibrium participation and price vectors, $\boldsymbol{a}^*, \boldsymbol{p^*}$, whatever the value of a_i^*. Hence, the statement in the Proposition holds.

Proof of Proposition 6

When $\sigma_i^2 = \sigma^2$ for all i, note that:

$$\phi_i(a_j = 0, \boldsymbol{a^*_{-j}}) - \phi_i(a_j = 1, \boldsymbol{a^*_{-j}}) = \sum_k [s_{ik}(a_j = 0, \boldsymbol{a^*_{-j}}) - s_{ik}(a_j = 1, \boldsymbol{a^*_{-j}})]\sigma^2.$$

By equation (5), then, the following equalities hold:

$$\phi_i(a_j = 0, \boldsymbol{a^*_{-j}}) - \phi_i(a_j = 1, \boldsymbol{a^*_{-j}}) = \sigma^2 \frac{\sum_t s_{ji}(a_j = 0, \boldsymbol{a^*_{-j}}) s_{it}(a_j = 0, \boldsymbol{a^*_{-j}})}{s_{jj}(a_j = 0, \boldsymbol{a^*_{-j}})} =$$
$$\sigma^2 \phi_i(a_j = 0, \boldsymbol{a^*_{-j}}) \frac{s_{ji}(a_j = 0, \boldsymbol{a^*_{-j}})}{s_{jj}(a_j = 0, \boldsymbol{a^*_{-j}})}.$$

By Theorem 5, $u_i(\boldsymbol{a}^*) = -\tau\phi_i(a_i = 0, \boldsymbol{a^*_{-i}})$, so Proposition 6 follows from the above equalities.

Proof of Theorem 6

Note first that by Lemma 2 (and, as is clear from equation (5) above), $\eta_{ij}(\boldsymbol{b})$ is weakly decreasing in $\boldsymbol{b}$, and so $\eta_{ik}(\mathbf{0}) \geq \eta_{ik}(\boldsymbol{b})$ for all $\boldsymbol{b} \geq \mathbf{0}$. Hence, if for all $\boldsymbol{b}$ such that $\boldsymbol{b}\mathbf{1}^{\boldsymbol{T}} = B$, $\eta_{ij}(\boldsymbol{b}) > \lfloor \frac{\zeta_k(\mathbf{0})}{\zeta_j(\boldsymbol{b})} \rfloor \eta_{ik}(\mathbf{0})$ then for any $\boldsymbol{b}'$ such that $\boldsymbol{b}'\mathbf{1}^{\boldsymbol{T}} \leq B$, $\eta_{ij}(\boldsymbol{b}') > \lfloor \frac{\zeta_k(\boldsymbol{b}')}{\zeta_j(\boldsymbol{b}')} \rfloor \eta_{ik}(\boldsymbol{b}')$.

It follows that for any $\boldsymbol{b}$ where $\hat{b}_k > 0$, there is always a profitable deviation for the social planner to some $\boldsymbol{b}'$ where $\hat{b}_k > \hat{b}'_k$ and $\hat{b}'_j > \hat{b}_j$. Hence, for any optimal $\boldsymbol{b}^*$, it must be that $\hat{b}_k = 0$.

As $B \geq c_i - \tau\phi(a_i = 0, \boldsymbol{a}^*)$, we know that $\hat{b}_l > 0$ for some group T_l, though, note that T_l may or may not be the group T_j. The inequality in the Theorem therefore holds.

Proof of Proposition 7

Lemma 7. *For any equilibrium,* $\boldsymbol{a}^*$, *for all* $i \in T_s$, $u_i(\boldsymbol{a}^*) = \hat{\tau}_s[\xi_i^2 - \phi_i(a_i = 0, \boldsymbol{a^*_{-i}})]$.

Proof. For any equilibrium, the service provider optimally sets a price, $p_i^* = c_i - \hat{\tau}_i\phi_i(a_i = 0, \boldsymbol{a}_{-i})$, for any agent who participates. If they did not, then there would always exist a profitable deviation such that the service provider could set a lower price $p_i^{'} < p_i^*$, which would still induce i to participate given $\boldsymbol{a}_{-i}$ and would yield a higher profit. As an agent, $i \in T_s$, who does not participate by definition receives $\hat{\tau}_s[\xi_i^2 - \phi_i(a_i = 0, \boldsymbol{a^*_{-i}})]$ the lemma holds. □

Now, consider some participation vector, $\boldsymbol{a}$. Given Lemma 7, $\hat{\tau}_i < 0$ and $\hat{\tau}_j > 0$, it follows that $u_i(\boldsymbol{a})$ is minimised at $\boldsymbol{a_{-i}} = \mathbf{0}$ and $u_j(\boldsymbol{a})$ is maximised at $\boldsymbol{a_{-j}} = \mathbf{0}$.

Note that $u_i(a_i = 0, \boldsymbol{a_{-i}} = \mathbf{0}) = u_j(a_j = 0, \boldsymbol{a_{-j}} = \mathbf{0})$, but, as $\phi_i(a_i = 0, \boldsymbol{a_{-i}})$ is weakly increasing in $\boldsymbol{a_{-i}}$ by Lemma 2, the statement in the Proposition holds.

Proof of Proposition 8

Suppose that, for some equilibrium with participation vector $\boldsymbol{a^*}$ and price vector $\boldsymbol{p^*}$, $\hat{\chi}_s(\boldsymbol{a}^*) \geq \hat{\chi}_t(\boldsymbol{a}^*)$. In this case, it must be that $\sum_{i\in T_s} a_i^* \geq \sum_{i\in T_t} a_i^*$. To see this, recall that as T_s and T_t are informationally identical then if $\sum_{i\in T_s} a_i^* = \sum_{i\in T_t} a_i^*$, $\hat{\chi}_s(\boldsymbol{a}^*) = \hat{\chi}_t(\boldsymbol{a}^*)$. Furthermore, this equality and group informativeness jointly imply that $\hat{\chi}_s(\boldsymbol{a}^*) > \hat{\chi}_t(\boldsymbol{a}^*)$ then $\sum_{i\in T_s} a_i^* > \sum_{i\in T_t} a_i^*$: as $\hat{\chi}_s(\boldsymbol{a}^*) = \hat{\chi}_t(\boldsymbol{a}^*)$ for all $\boldsymbol{a}_{-t,s}$, it follows that if $\sum_{i\in T_s} a_i^* < \sum_{i\in T_t} a_i^*$, then $\hat{\chi}_s(\boldsymbol{a}^*) < \hat{\chi}_t(\boldsymbol{a}^*)$.

Note that for any equilibrium price vector $\boldsymbol{p^*}$, it must be the case that $p_i^* = c_i - \tau\chi_i(a_i = 0, \boldsymbol{a_{-i}})$, for an analogous reason that statement 1 in Theorem 1 holds: if $p_i^* < c_i - \tau\chi_i(a_i = 0, \boldsymbol{a_{-i}})$ when $a_i^* = 1$ then i would prefer to deviate and not participate and if $p_i^* > c_i - \tau\chi_i(a_i = 0, \boldsymbol{a_{-i}})$, then the service provider can still induce i to deviate with some price $p_i^{'}$, where $c_i - \tau\chi_i(a_i = 0, \boldsymbol{a_{-i}}) \leq p_i^{'} < p_i^*$.

Now, consider an alternative equilibrium activity-price pair $\boldsymbol{a}^{'}$, $\boldsymbol{p}^{'}$, which are identical to $\boldsymbol{a^*}$and $\boldsymbol{p^*}$ except there exist a pair $i \in T_s$ and $j \in T_t$ such that $a_i^* = a_j^{'} = 1$ and $a_i^{'} = a_j^* = 0$ with prices such that i and j are just indifferent between participating and not participating in the potential equilibrium they participate, and 0 otherwise. [2]

As $c(y_s) > c(y_t)$ and $\chi_i(a_i = 0, \boldsymbol{a_{-i}}) > \chi_j(a_j = 0, \boldsymbol{a_{-j}})$, it follows that $\sum_{i\in A_1} p_i^{'} < \sum_{i\in A_1} p_i^*$. If $\sum_{i\in T_s} a_i^* = \sum_{i\in T_t} a_i^*$, then by the informationally identical property, it must be that $\chi(a^*) = \chi_j(a^{'})$, and so $\boldsymbol{a^*}$, $\boldsymbol{p^*}$ cannot be part of an equilibrium in this case.

[2]If there exists no $j \in T_t$ such that $a_j^* = 0$, then it must be that every T_s agent participates if $\hat{\chi}_s(\boldsymbol{a}^*) \leq \hat{\chi}_t(\boldsymbol{a}^*)$, and so the statement in the Proposition automatically holds.

Group submodularity implies that, starting at $\sum_{i \in T_s} a_i = \sum_{i \in T_t} a_i$ and for any $\boldsymbol{a}_{-s,t}$, an increase in the number of active T_s group agents reduces $\gamma_i(\boldsymbol{a}_{-i})$ more than $\gamma_j(\boldsymbol{a}_{-j})$ for $i \in T_s$ and $j \in T_t$, which then implies when $\sum_{i \in T_s} a_i^* > \sum_{i \in T_t} a_i^*$, $\chi(a^*) < \chi_j(a^{'})$, and so, again, $\boldsymbol{a}^*$, $\boldsymbol{p}^*$ cannot be part of an equilibrium.

Proof of Proposition 9

Result 2 follows from the fact that for any potential equilibrium in which $a_j^i = 1$ and $p_j^i > c_j - \tau\phi_j(\boldsymbol{a_j} = \boldsymbol{0}, \boldsymbol{a}_{-j})$, intermediary i always has an incentive to deviate and set some price $c_j - \tau\phi_j(\boldsymbol{a_j} = \boldsymbol{0}, \boldsymbol{a}_{-j}) \leq p^{'} < p_j^i$, receiving a higher profit while still inducing a_i to participate.

To establish result 1 in the Proposition, it is necessary to establish the following Lemma:

Lemma 8. *For any equilibrium participation matrix, $\boldsymbol{a}$, if $a_j^i = 1$, then $a_j^k = 0$ for all $k \neq j$.*

Proof. Suppose instead that for some equilibrium participation matrix, $\boldsymbol{a}$, if $a_j^i = a_j^k = 1$, for some $k \neq j$. By the same argument as the proof of Theorem 1, $p_j^s \geq c_j - \tau\phi_j(\boldsymbol{a_j} = \boldsymbol{0}, \boldsymbol{a}_{-j})$ for any s where a_j^s. Suppose, without loss of generality, $p_j^i = p_j^k = c_j - \tau\phi_j(\boldsymbol{a_j} = \boldsymbol{0}, \boldsymbol{a}_{-j})$ for two intermediaries, i and k, with $p_j^s = 0$ for all other intermediaries.

Consider the equilibrium prices for j's data when it is sold to the service provider, q_j^i and q_j^k. The products the two intermediaries are selling are homogeneous, and thus by standard arguments the only set of equilibrium price for j's data is $q_j^i = q_j^k = 0$.

It follows that, in equilibrium, if $a_j^i = 1$ for some i, then an intermediary $k \neq i$ sets a price $p_j^k < c_j - \tau\phi_j(\boldsymbol{a_j} = \boldsymbol{0}, \boldsymbol{a}_{-j})$ and so $a_j^k = 0$. □

Given Lemma 8, it follows that only a single intermediary (at most) holds data on a given agent j. As established in Theorem 1, $v_j(\boldsymbol{\alpha}^*) = \theta_i(\boldsymbol{\alpha}^{-j})$. Hence, if in some potential equilibrium an intermediary i sets $q_j^i < \theta_j(\boldsymbol{\alpha}^{-j})$,there is always a deviation in which i sets a price $q^{'}$ $q_j^i < q^{'} \leq \theta_j(\boldsymbol{\alpha}^{-j})$ which yields a higher profit for i. Result 3 in the Proposition follows directly from this analysis.