INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Competitive equilibrium based personal data market model

J Ujwala Rekha[1]∗**, K Shahu Chatrapati**[2]

1 Associate Professor of CSE, JNTUH College of Engineering Hyderabad, Telangana, India
2 Professor of CSE, JNTUH College of Engineering Manthani, Telangana, India

Abstract

Background/Objectives: In a digital economy, with the increasing commercial value, personal data is viewed as a commodity to be bought and sold. Data owners expect an appropriate compensation for trading off their privacy depending on how they value their privacy. Simultaneously, data consumers want to maximize their utility which is dependent on the value derived from the data. Consequently, a data market model that optimally recompenses data owners and maximizes the profits of data consumers is required. **Methods:** In this study, a data market model and pricing mechanism based on Fisher market model and competitive equilibrium are presented where the value derived from the data is calculated from information entropy. The proposed data model and the pricing mechanism jointly and simultaneously maximize the profit of data owners and the utility of data buyers. **Findings:** Experiments are conducted on adult data set to validate the efficacy of the proposed approach. Data owners are classified as risk averse, risk neutral, risk taking and privacy regarders. Subsequently, prices of data samples are adjusted to reach equilibrium as defined by the Fisher market model maximizing simultaneously and jointly the profit of data owners and the utility of data buyers. **Applications:** The proposed competitive equilibrium based personal data market model can be used to find equilibrium prices and bundles of data samples for each data buyer at these prices maximizing the utility of data buyer subject to his budgetary constraints and data requirements, and data owners' privacy preferences.

Keywords: Personal data; Competitive equilibrium; Fisher market; Information entropy; Data buyers; Data owners; Data market

1 Introduction

Personal data is continually being generated, gathered and shared at an unprecedented level which is useful in improving the products and services. For instance, personal information is useful in a variety of domains which include medical research, crime analysis, web usage analysis, customer behavior analysis and risk analysis. However, a raw personal data may consist of information about individuals that is sensitive. Consequently, use of personal data is constrained owing to privacy concerns. Due to the mismatch in demand and supply, the concept of personal data market has emerged where personal data is viewed as a commodity. Numerous studies have emphasized that

a monetary compensation ought to be awarded to the data owners for trading off their privacy[1–4]. However, evaluating the costs and benefits of preserving or disseminating personal data and quantifying the worth of personal data is notably challenging.

There are three stakeholders involved in the personal data market:

1. The *data owners* who make available their personal information to data market along with their privacy preferences such as when, how and for what purpose his/her personal data can be collected and shared to others. Furthermore, a data owner obtains an appropriate monetary compensation in return for trading off his privacy.
2. The *data buyers* who seek data from the data market for either research or commercial purpose. A data buyer makes payment to the data market in proportion to the amount of information acquired and/or value derived from the information obtained.
3. The *data market* is a trusted intermediary between data owners and data buyers. A data market gathers personal data of data owners together with their privacy preferences and disseminates data to data buyers, meeting the obligation of preserving privacy choices.

Numerous studies exist on personal data market from the perspective of data privacy and can be classified into information-technology based and economic-based approaches. Technology-based approaches protect privacy while sharing personal data by masking, generalization and suppression techniques for anonymizing the personally identifiable attributes in the data[5–8]. While technology-based approaches can be used in data mining applications to find aggregate or statistical information or to obtain interesting patterns in the data, they are not suitable in applications such as crime analysis and marketing where sensitive information and individual identities are required.

Economics-based approaches rely on pricing mechanism for privacy compensation. In these approaches, personal data is considered as a commodity and individuals trade off their privacy for monetary gain[9,10]. The worth of the personal data is determined using economic tools. Laudon[11] advocates a market-based approach where individuals have the control over the information about them, and thus receive a fair compensation for trading off their privacy. Gkatzelis et al. introduces a market that makes payments to the individuals according to their privacy choices[12]. At the same time, Gkatzelis et al. introduces a mechanism that elicits individuals to truthfully report their privacy preferences. Similarly,[13] proposes a fair privacy compensation mechanism that depends on the privacy attitude of the individuals and the sensitivity of the information provided by each individual. In another study[14], a big data market model is introduced and an optimal pricing scheme is formulated as a Stackelberg game to maximize the profit of the data source. Yet another study[15], presents an optimal pricing scheme based on the quality of the provided data. However, studies[14] and[15] do not consider the privacy preferences of the data providers.

Furthermore, there are studies that integrate technology-based approaches with economic-based models that simultaneously preserve individuals' privacy and increases the value of personal information. For instance in[16], a mechanism is proposed that integrates the technology based security mechanism with the market model that compensates individuals according to their privacy attributes. However, the proposed mechanism of[16] has limited applicability since the economic model employed is dependent on the specific security mechanism.

The key determinants in pricing the personal data include privacy attitudes of the data owners and the value derived by the data buyers from the information obtained. Privacy is the right of an individual and each individual's privacy attitudes are different. Consequently, each person may be willing to sell his personal data for different prices. For instance, individuals less concerned about privacy may intend to sell their personal data for a small payment while others who are more concerned about privacy may consent to sell their personal data for higher payments. Furthermore, there may be some individuals who pretend to regard privacy as more valuable and misreport the cost of their personal data in an expectation of receiving higher payments. On the other hand, data buyers desire to obtain maximum utility from the information acquired at minimum subscription cost. For instance, a data buyer may desire to obtain those samples of data that yield more information content. Thus both data owners and data buyers are selfish and each of them intend to maximize their own objectives.

In this study, a pricing mechanism based on competitive equilibrium theory is presented which jointly optimizes privacy loss of data owners and utility of data buyers. In this study, a data buyer desires to buy the samples that maximize the information entropy.

The remaining of the study is organized as follows: While Section 2 presents an overview of competitive equilibrium theory Section 3 describes calculation of information entropy. Section 4 describes the proposed pricing model and Section 5 demonstrates experimental results. Finally, conclusions are drawn in Section 6.

2 Competitive Equilibrium

Competitive equilibrium also termed as Walrasian equilibrium is a market state which satisfies the following conditions:

1. Producers/sellers and the consumers/buyers arrive at an equilibrium price for a product.
2. At an equilibrium price, the market clears; in other words demand equals supply for a product.
3. At equilibrium price, sellers maximize their profits and buyers maximize their utility subject to resource and technological constraints.

In this study, we will discuss the competitive equilibrium of a Fisher market model that contains m commodities and n buyers. Each buyer i is endowed with revenue $r_i > 0$ and utility function $u_i : [0,1]^m \to R$ that determines the choice for consuming various commodities. Similarly, each commodity j has pre-specified quantity c_j. Each buyer spends his revenue to consume a bundle of goods such that his utility function is maximized. Let $X_i = (x_{i1}, \ldots, x_{im}) \in [0,1]^m$ represents the bundle of goods consumed by buyer i; where $x_{ij} = 1$ implies good j is consumed and $x_{ij} = 0$ implies good j is not consumed by buyer i. Each commodity j has a price p_j determined by an auction that will be described below. A competitive equilibrium is a vector of prices $P = (p_1, \ldots, p_m) \to R_+^m$ and bundle of goods X_i for $1 \le i \le n$ such that the market clears and the utility of each buyer is maximized[17]. Formally, the conditions for the market can be stated as follows:

1. For all $1 \le i \le n$, the vector X_i is the maximizer of utility of buyer i, that is $\max_{X_i} u_i(X_i)$ subject to the constraint $\sum_{k=1}^m p_k * x_{ik} \le r_i$.
2. For each commodity j, $\sum_{k=1}^n x_{kj} = c_j$.

Walras introduced a trial and error process called tâtonment process for determining equilibrium prices[18]. Each buyer reports his demands for goods maximizing his utility at prices announced by an auctioneer. The market is in disequilibrium if there is either an excess demand or supply. Buying and selling does not take place at disequilibrium instead the auctioneer either lowers or raises the prices of goods depending on demand and supply of corresponding goods. The process terminates if the market reaches an equilibrium that is when the market clears. In other words, equilibrium is reached if demand and supply of goods is equal. According to[19], equilibrium is guaranteed if each commodity is consumed by at least one consumer and each consumer buys at least one item. More formally, Arrow and Debreu prove that the market converges to equilibrium if the utility functions of the agents are concave[20].

3 Information Entropy

Information entropy introduced by Claude Shannon is a measure of 'uncertainty' or 'surprise' associated with a random variable. Let X be a discrete random variable whose possible outcomes are $x_1, \ldots, x_n$, then the entropy of variable X is defined as

$$
\begin{aligned}
H(X) &= \sum_{i=1}^n p(x_i) log_2 \frac{1}{p(x_i)} \\
&= -\sum_{i=1}^n p(x_i) log_2 p(x_i)
\end{aligned}
\tag{1}
$$

where $p(x_i) = Pr(X = x_i)$ is the probability of ith outcome of X.

Suppose a dataset contains n samples each with k attributes, then the entropy of ith sample can be calculated as follows:

$$
H(x_i) = -\sum_{j=1}^k p(x_{ij}) log_2 p(x_{ij})
\tag{2}
$$

The total entropy of all samples in a dataset D is computed as follows:

$$
\begin{aligned}
H(D) &= \sum_{i=1}^n H(x_i) \\
&= -\sum_{i=1}^n \sum_{j=1}^k p(x_{ij}) \log_2 p(x_{ij})
\end{aligned}
\tag{3}
$$

A skewed probability distribution is unsurprising and has low entropy and lesser information content. On the other hand, a balanced probability distribution is surprising and has high entropy and more information content. The contribution of information content of ith sample in the whole dataset can be calculated according to[21] as follows:

$$
IC_i = \frac{H(x_i)}{H(D)}
\tag{4}
$$

It can be noted that $\sum_{i=1}^n \frac{H(x_i)}{H(D)} = 1$.

According to[22] the information entropy is concave, continuous and continuously differentiable function of x, where x is a compact convex set.   (2)

4 Proposed Pricing Model

We consider a data market D consisting of n data owners and m data buyers. Let x_i denote the sample corresponding to data owner i and $X_j \subseteq D$ denote the bundle of samples data buyer j has access to. The utility of a data buyer j is derived from the information entropy of bundle X_j and is computed as follows:

$$u_j(X_j) = \sum_{x_i \in X_j} H(x_i) \qquad (5)$$

where $H(x_i)$ is the information entropy of data sample x_i defined over the sample set $X_j \subseteq D$ calculated according to (2).

While each data owner i reports his initial price p_i for which he is willing to sell his personal data along with the privacy preferences such as when, how and for what purpose his personal data can be shared to others, each data buyer j reports his budget b_j and data requirements to the data market. Data requirements may specify the number of data samples and/or specific kind of data samples required. Data market bundles the samples to each data buyer at given prices maximizing the utility of the buyer subject to his budget constraints and data requirements according to the privacy preferences of each data owner.

In order to obtain market equilibrium, the data market adjusts the prices of each data owner's sample depending on demand and supply, and the information entropy of the corresponding data owner's sample. If p_i is too high, there is possibility of x_i not being sampled and bought by any data buyer. On the other hand, if p_i is too low, the data owner of x_i receives less payment for trading off his privacy. Therefore, in addition to the initial price p_i, the data market elicits maximum price $pmax_i$ and minimum price $pmin_i$ at which each data owner is willing to sell his personal data. A risk taker, reports maximum and minimum prices such that $pmin_i = p_i < pmax_i$; a person who is risk averse reports maximum and minimum prices such that $pmin_i < p_i = pmax_i$; and a person who is risk neutral, reports maximum and minimum prices such that $pmin_i < p_i < pmax_i$. In addition to risk takers and persons who are risk averse and risk neutral, there are privacy regarders who value privacy more than incentives and such a person reports maximum and minimum prices such that $pmin_i = p_i = pmax_i$, where price p_i is generally very high. A sample has high probability of being sampled if it has more information entropy and/or lesser price, and vice versa. Therefore, data market increases the price of sample that has more information entropy and lesser price to increase the profit of the corresponding owner. On the other hand, the price of the sample is lowered if it has less information entropy and higher price to increase its probability of being sampled. However, data market adjusts the prices of the samples within the range of maximum and minimum prices specified by the data owners.

Equilibrium is reached if a vector of prices $P^* = (p_1^*, \ldots, p_n^*) \in R_+^n$ and bundle of samples X_j^* is determined such that the market clears and the utility of each data buyer is maximized subject to his budget and data constraints. Formally, the conditions for equilibrium in the context of data market can be defined as follows:

1. For each data buyer j, X_j^* is the maximizer of utility function $\max_{X_j^*} u_j\left(X_j^*\right)$ at given prices P^* and budget b_j and

2. The market clears, that is $\sum_{j=1}^{m} \sum_{x_i \in X_j^*} p_i^* = \sum_{j=1}^{m} b_j$

Details of computing competitive equilibrium prices and bundles of samples for each buyer at equilibrium prices maximizing his utility subject to his budget constraints and data requirements and privacy preferences of data owners is presented in Algorithm 1. Since, information entropy is a concave, strictly continuous and continuously differentiable function of x where x is a compact convex set; the utility functions of the data buyers $u_j(X_j)$ for $1 \le j \le m$ are concave, strictly continuous and continuously differentiable functions of X_j where X_j is a compact convex set. According to [20], competitive equilibrium exists for the data market and pricing model defined above since the necessary and sufficient conditions for the existence of competitive equilibrium are fulfilled.

5 Experiments and Results

In order to substantiate the efficacy of the pricing mechanism presented in this study, experiments are conducted on Adult dataset of UCI machine learning repository. The dataset consists of 48,842 instances described by 14 attributes. However, random samples of 20,000 are drawn from 48,842 samples and 7 attributers are considered in the experiments. The description of the attributes is presented in Table 1.

In the experiments, 120 buyers are considered each of them requesting 8000 samples. For simplicity, budget of all buyers is set to 1. The initial price p_i, maximum price $pmax_i$ and minimum price $pmin_i$ are set in the range [0, 1] for each sample such that the number of persons who are risk taking, risk averse, risk neutral and privacy regarders is 5000.

In Figure 1, the average number of records sampled at minimum price, initial price and maximum price are plotted for samples whose information entropy is 1, 0.5 and 0.2. It can be observed that the average number of records sampled is decreasing

Algorithm 1: Computation of competitive equilibrium prices

Input

        Data Samples: $D = x_1, .., x_n$

        Initial Prices of Samples: $p_1, .., p_n$

        Maximum Prices of Samples: $pmax_1, .., pmax_n$

        Minimum Prices of Samples: $pmin_1, .., pmin_n$

        Budgets of buyers:$b_1, .., b_m$

Output

        Equilibrium Prices: $p_1^*, .., p_n^*$

        Bundles of Samples: $X_1, .., X_m \; where \; X_j \subseteq D$

1. $p_i^* \leftarrow p_i \; for \; i = 1 \, to \, n$

2.   **Loop**

        2.1.   At prices $p_1^*, .., p_n^*$ determine bundles of samples $X_1, .., X_m$ such that each data buyer's utility function (5) is maximized subject to budget and data constraints in accordance with privacy preferences of data sellers

        2.2.   **If** $\sum_{j=1}^m b_j - \sum_{j=1}^m \sum_{x_i \in X_j^*} p_i^* \geq \alpha$ **then**

                Adjust prices $p_1^*, .., p_n^*$ in accordance with information entropy and demand of the corresponding sample within the range of maximum and minimum prices specified by the data owners.

          Else

               Exit Loop

3.   Report equilibrium prices and bundles of samples $X_1, .., X_m$ at these prices

Table 1. Characteristics of Attributes

Attribute	Cardinality	Sensitivity
Age	65	No
Gender	2	No
Marital Status	3	No
Race	4	No
Native Country	25	No
Education	10	No
Occupation	10	Yes

in all three cases. For samples whose information entropy is 1, average number of records sampled does not vary much as we move from $pmin_i$ to p_i; however, there is a sharp drop as we move from p_i to $pmax_i$. For samples whose information entropy is 0.2 there is a sharp decline from $pmin_i$ to $pmax_i$.
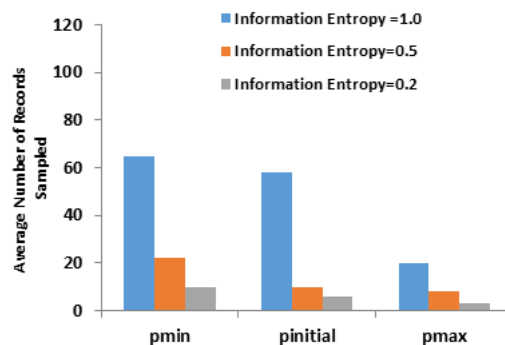

Fig 1. Average number of records sampled at various prices

Figure 2 presents the average number of records sampled at equilibrium prices for persons who are risk averse, risk neutral, risk taking and privacy regarders. As expected, average number of records sampled for privacy regarders is the least and highest for persons who are risk averse. Furthermore, the average number of records sampled for persons who are risk neutral is less than risk aversers but greater than risk takers who in turn are sampled more than privacy regarders.
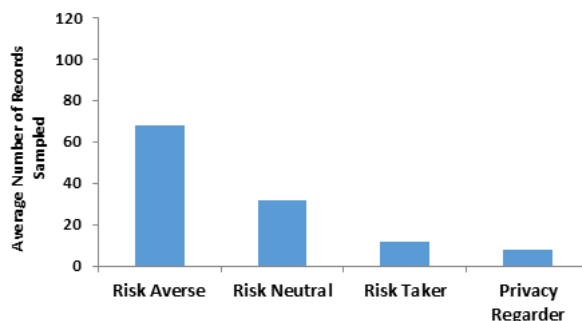

Fig 2. Average number of records sampled at equilibrium prices

Figure 3 illustrates average profits of the persons who are risk averse, risk neutral, risk taking and privacy regarders at equilibrium. It can be noted that the variations in the average profits are not so significant and almost same for persons who are risk averse, risk neutral and risk taking but very much less for privacy regarders because of rigid restrictions on their privacy.
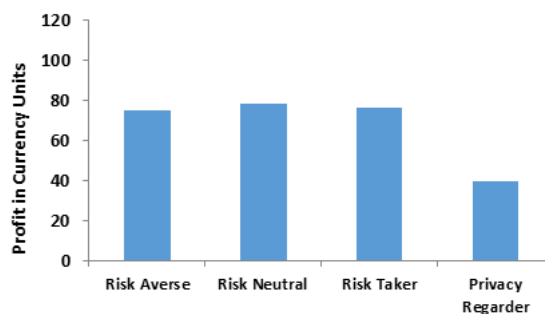

Fig 3. Profit earned by different kinds of data owners

Figure 4 exhibits plots of the equilibrium prices of samples versus information entropy. As prices are adjusted in accordance with demand which in turn is dependent on information entropy, the prices are linearly increasing with the increase in information entropy.
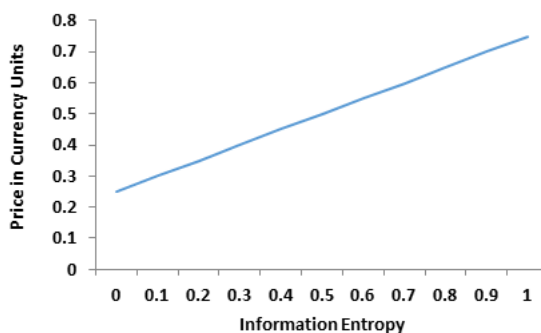

Fig 4. Profit Vs. information entropy at equilibrium prices

Figure 5 shows the variations of average equilibrium prices for persons who are risk averse, risk neutral and risk taking. While the average equilibrium price of risk takers is the highest, it is the least for risk aversers.
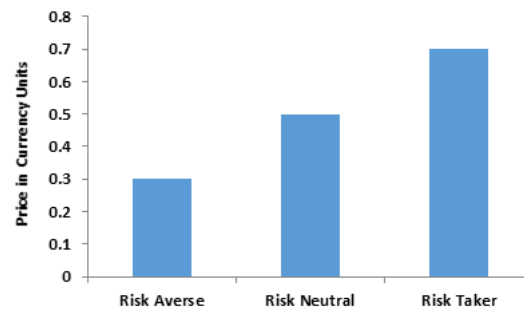

Fig 5. Equilibrium prices of various data owners

Figure 6 plots utility against buyers at equilibrium prices. In order to illustrate, the experiments are conducted on 10 buyers instead of 120. Since the budget of all buyers is equal, the utility of all the buyers is almost same with little variations since competitive equilibrium maximizes the objective of all the agents simultaneously.

Fig 6. Utility gained by various data buyers

6 Conclusions

In a data market, the objective of the data owners is to ensure for an appropriate compensation in lieu of trading off their privacy. At the same time, the objective of the data buyers is to derive maximum utility from the data obtained. The value derived from the data is primarily dependent on the information entropy of the data; if information entropy is more, then the utility derived is more and vice a versa. Hence, in this study, a personal data market model and pricing mechanism based on competitive equilibrium is presented that simultaneously and jointly optimizes the profit of data owners and the utility of data buyers. Experiments validate the propositions and the effectiveness of the proposed approach. The proposed approach is applicable for a static data market. In the future, a solution to a dynamic market based on competitive equilibrium should be studied.

References

1) Ghosh A, Roth A. Selling privacy at auction. *Proceedings of the 12th ACM conference on Electronic commerce*. 2011;p. 199–208.
2) Li C, Li DY, Miklau G, Suciu D. A theory of pricing private data. *Communications of the ACM*. 2017;60(12):79–86. doi:10.1145/3139457.
3) Riederer C, Erramilli V, Chaintreau A, Krishnamurthy B, Rodriguez P. For sale: your data: by you. *Proceedings of the 10th ACM WORKSHOP on Hot Topics in Networks*. 2011;p. 1–6.
4) Bergemann D, Bonatti A, Gan T. 2020. doi:10.2139/ssrn.3548336.
5) Nussbaum E, Segal M. 2019. Available from: https://arxiv.org/abs/1905.11694.
6) Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. *In21st International conference on data engineering (ICDE'05*. 2005;p. 217–228.
7) Nayahi JJV, Kavitha V. Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop. *Future Generation Computer Systems*. 2017;74:393–408. doi:10.1016/j.future.2016.10.022.
8) Man N, Li X, Wang K. A Privacy Protection Model Based On K-Anonymity. *In2017 International Conference Advanced Engineering and Technology Research*. 2017.

9) Bataineh AS, Mizouni R, Bentahar J, Barachi E, M. Toward monetizing personal data: A two-sided market analysis. *Future Generation Computer Systems*. 2019. doi:10.1016/j.future.2019.11.009.

10) Choi JP, Jeon DS, Kim BC. Privacy and personal data collection with information externalities. *Journal of Public Economics*. 2019;173:113–137. doi:10.1016/j.jpubeco.2019.02.001.

11) Laudon KC. Markets and privacy. *Communications of the ACM*. 1996;39(9):92–104. doi:10.1145/234215.234476.

12) Gkatzelis V, Aperjis C, Huberman BA. Pricing private data. *Electronic Markets*. 2015;25:109–123. doi:10.1007/s12525-015-0188-8.

13) Yang J, Xing C. Personal Data Market Optimization Pricing Model Based on Privacy Level. MDPI AG. 2019. doi:10.3390/info10040123.

14) Niyato D, Alsheikh MA, Wang P, Kim DI, Han Z. Market model and optimal pricing scheme of big data and Internet of Things (IoT). *In2016 IEEE International Conference on Communications (ICC)*. 2016;p. 1–6.

15) Yang J, Zhao C, Xing C. Big Data Market Optimization Pricing Model Based on Data Quality. *Complexity*. 2019;2019:1–10. doi:10.1155/2019/5964068.

16) Garfinkel R, Gopal RD, Nunez M, Rice DO. Secure electronic markets for private information. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*. 2006;36(3):461–471. doi:10.1109/tsmca.2006.871794.

17) Nisan N, Roughgarden T, Tardos E, Vazirani VV, et al., editors. Algorithmic Game Theory. Cambridge University Press. 2007. Available from: https://staff.fnwi.uva.nl/s.botan/algorithmic-game-theory.pdf.

18) Walras L. Elements of pure economics. *Routledge*. 2013. Available from: https://pdfs.semanticscholar.org/38b9/e65f8eaacce89dcbc73d1292c9d73113f907.pdf.

19) Maxfield RR. General equilibrium and the theory of directed graphs. *Journal of Mathematical Economics*. 1997;27(1):23–51. doi:10.1016/0304-4068(95)00763-6.

20) Arrow KJ, Debreu G. Existence of an Equilibrium for a Competitive Economy. *Econometrica*. 1954;22(3):265–265. doi:10.2307/1907353.

21) Shen Y, Guo B, Shen Y, Duan X, Dong X, Zhang H. A pricing model for big personal data. *Tsinghua Science and Technology*. 2016;21:482–90. doi:10.1109/TST.2016.7590317.

22) Entropy, convexity, and matrix scaling. . Available from: https://homes.cs.washington.edu/~jrl/teaching/cse599swi16/notes/lecture1.pdf.