HHS Public Access

Author manuscript

Decis Sci. Author manuscript; available in PMC 2021 November 02.

Published in final edited form as:

Decis Sci. 2021 April ; 52(2): 393–426. doi:10.1111/deci.12442.

Valuing Personal Data with Privacy Consideration

Xiao-Bai Li[†],

Department of Operations and Information Systems, Manning School of Business, University of Massachusetts Lowell, One University Ave., Lowell, MA 01854

Xiaoping Liu,

D'Amore-McKim School of Business, Northeastern University, 360 Huntington Ave., Boston, MA 02115

Luvai Motiwalla

Department of Operations and Information Systems, Manning School of Business, University of Massachusetts Lowell, One University Ave., Lowell, MA 01854

Abstract

A key challenge in information privacy research is how to value personal data with privacy consideration. In this study, we propose an experimental auction approach for valuing personal data. We use the generalized second-price auction to assess the monetary values of individuals' identity, demographic, and private information. We find that individuals' economic valuation of personal data is consistent with their actual self-disclosure behaviors. The economic valuation approach also produces results that are consistent with some well-accepted observations about consumer demographics and privacy. On the other hand, individuals' stated privacy preferences and attitudes are not consistent with their economic valuation. The findings of this study suggest that the proposed approach can be an effective mechanism for measuring personal data privacy. This study also provides important insights into valuing personal information for practical uses with several implications to policy decision makers, corporate executives and managers, data analysts, as well as decision science researchers.

Subject Areas:

Auctions; Data Science; Measurement; Personal Data; Valuation

INTRODUCTION

How much is an individual's personal data worth? In June 2013, *Financial Times* launched an interactive personal data calculator developed based on data-broker industry's pricing scheme. Using the calculator, the basic nonidentifiable demographic data such as age, gender, and race of a person are worth less than a penny. Even after including additional information, a normal person's data are still worth less than a dollar (Steel, Locke, Cadman, & Freese, 2013). On the other hand, a *Huffington Post* article reported that a federal judge

[†] Corresponding author. Xiaobai_Li@uml.edu.

claimed publicly that he would be willing to pay $2,400 per year to protect his family's online privacy (Sledge, 2013). A small data-broker company called Datacoup has paid its customers $8 a month in exchange for accessing the customers' social media accounts, such as Facebook and Twitter, and their credit-card and debit-card transaction records (Simonite, 2014). Datacoup's business plan is to make money by charging companies for the aggregated information and analysis results obtained from its customers' data, without disclosing the customers' personally identifying information (Datacoup, 2019). According to an industry estimate, an individual's personal data would be worth between half a cent and $1,200 (Madrigal, 2012). In January 2018, two U.S. Senators, Elizabeth Warren and Mark Warner, introduced a bill that would require a credit reporting firm such as Equifax and Experian to pay $100 to each individual whose personal information is stolen in a data breach and $50 for each additional piece of personal information compromised (Warren & Warner, 2018). These snippets provide a very mixed picture on the economic value of personal data, opening up an opportunity for researchers to explore new approaches for measuring the value of personal data in the context of consumer privacy.

The economic value of personal data is of considerable interest to policy decision makers, decision science and information systems researchers, and practitioners in this data-driven business environment. Laudon (1996) introduced the idea of a regulated National Information Market (NIM) that could allow personal information to be bought and sold like a commodity. In NIM, individuals would decide whether or how much their personal information can be released for secondary use, and data collectors and users would pay for the acquisition and use of this information. This idea, however, has not been put into practice thus far. Instead, the business of buying, aggregating, and selling consumer data has become a multibillion-dollar industry. Major players such as Acxiom and Datalogix actually do not interact with individual consumers; instead, they acquire consumer data from organizations that own large collections of customer data (FTC, 2014). There are also small startups (e.g., Datacoup, Handshake, and Meeco) that serve as intermediaries between individual consumers who are willing to accept payment in exchange for their personal data and companies that are interested in purchasing these data (Palet, 2014). A sound approach to value personal data will be useful for government regulators to establish effective privacy policies and for businesses in the data economy to be transparent about their customer data collection and sharing practices.

From a privacy research perspective, measuring the value of personal data is an important topic of research in multiple disciplines, including marketing, management, information systems, and decision science (Syed, Dhillon, & Merrick, 2019). To scientifically study the relationships between information privacy and other observable or latent constructs, it is necessary to measure and quantify the value of personal information. In some privacy research domains that are closely related to the advances in information technology, measuring the value of personal information is an essential part of the research objectives. These research areas will benefit greatly if a mechanism exists to comprehensively measure the economic value of personal data. An example of such an active research topic is *privacy calculus*, which asserts that personal data have value and thus considers privacy-related decision-making as a rational process where individuals assess or measure the expected

costs/risks of disclosing personal data against the potential benefits (Phelps, Nowak, & Ferrell, 2000; Culnan & Bies, 2003; Dinev & Hart, 2006).

Privacy calculus measures the trade-off between the costs/risks and the benefits of disclosing and sharing personal information (Hui & Png, 2006). Although efforts have been made to measure and value personal privacy, the issue remains to be one of the most challenging problems in information privacy research (Bélanger & Crossler, 2011; Pavlou, 2011). According to a review of information privacy literature (Smith, Dinev, & Xu, 2011), "almost all empirical privacy research in the social sciences relies on measurement of a privacy-related proxy of some sort." Most often, measurements of privacy concern and attitude are used as the proxy (Smith et al., 2011). Privacy measurement scales widely used in research and practice include Westin privacy indexes (Kumaraguru & Cranor, 2005), the concern for information privacy (Smith, Milberg, & Burke, 1996), and Internet users' information privacy concerns (IUIPC) (Malhotra, Kim, & Agarwal, 2004), among others. These measurement scales rely on survey responses to quantify and assess privacy. However, studies have found that individuals' stated privacy concerns and attitudes do not necessarily match up to their actual behaviors in disclosing private information. Thus, it is important to have a mechanism that better characterizes and reveals individuals' actual privacy cost/benefit concerns in the context of privacy calculus.

From an economics standpoint, personal information is considered as a commodity with economic values that can be measured in monetary terms (Laudon, 1996). More specifically, personal information is viewed as a nonmarket commodity as consumers generally do not trade their personal information in markets, i.e., in a direct or first exchange setup (Culnan & Bies, 2003). To assess economic value of nonmarket commodities, researchers typically conduct an experimental economics study that facilitates a setting where a direct or first exchange of nonmarket commodities (e.g., personal data) and monetary payoff takes place (List & Gallet, 2001). Individuals' economic valuation of their personal data in this exchange setting is considered a more valid and reliable measure of their privacy dispositions than their stated privacy preferences (List & Gallet, 2001). This work responds to the challenging issue of determining consumers' real privacy concerns and examines whether a direct exchange approach for economic valuation of personal data is a better proxy for measuring their information privacy concerns than the survey-response-based methods.

The research question of this study is how to measure economic value of personal data with privacy consideration. The primary objective of this study is to develop an economic valuation mechanism to measure the value of personal data. Our study explores an auction mechanism to measure the value of personal data with privacy consideration. As a secondary objective, this research also examines if the results of our auction-based approach are consistent with the empirical findings on the relationships between economic valuation of personal data and a set of variables representing individuals' self-disclosure behaviors, their stated privacy preferences, and their demographic characteristics. Our work contributes to information privacy research in several aspects. First, we propose using an auction mechanism to elicit individuals' valuation of their personal data in an economic exchange. The specific mechanism is the generalized second-price (GSP) auction, which is easy

to implement and effective in determining a large number of winners simultaneously, overcoming the limitations of the auction mechanisms used in prior studies. We use the GSP auction to acquire personal data and assess the value of the identity, demographic, and private information. Second, we find that individuals' valuation of personal data is consistent with their actual self-disclosure behaviors. The valuation approach also produces results that are consistent with some well-accepted observations about individual demographics and privacy. On the other hand, individuals' stated privacy preferences and attitudes are not consistent with their economic valuation of personal data. Finally, our valuation model provides important insights into valuing personal information for practical uses with several implications to policy decision makers, corporate executives and managers, data analysts, and decision science and privacy researchers.

The rest of the article is organized as follows. The next section reviews related work on consumer privacy with a focus on economic valuation of personal information. Following that, we present our research framework, including our model for valuing personal information. We then describe our research methodology and procedure. Subsequently, we provide the results of data analysis, which is followed by some in-depth discussions. The final section concludes our findings and provides future research directions.

RELATED WORK

Individuals' valuation of their personal data depends on the ownership of the data. If the person owns the data, the appropriate value measure is her willingness to accept (WTA) compensation to sell her data; otherwise, the appropriate value measure is her willingness to pay (WTP) to protect her information (Horowitz & McConnell, 2002; Lusk & Shogren, 2007, p. 35). Prior empirical studies have found that WTA value is usually higher than WTP value (Horowitz & McConnell, 2002). Research on economic valuation of personal data has focused largely on the WTA case (as we describe in this section) because WTP more likely depends on a specific context (e.g., the dependability of the data owner, the security and accessibility of the data to a third party, etc.). Our study also focuses on the WTA scenario.

An early attempt to quantify monetary value of personal data was undertaken in Hann, Hui, Lee, and Png (2002, 2007). One of the important findings of their research is that individuals are generally willing to accept economic benefits for sharing their personal information. Using conjoint analysis on survey data, the authors find that, for the U.S. participants, protection against errors, improper access, and secondary use of personal information is worth between $30.49 and $44.62. A critical limitation of their study was that it did not involve actual exchange of personal data for a payment. Essentially, the dollar amount found by the study is a "stated" value of personal data, rather than the actual value. Another study (Morey, Forbath, & Schoop, 2015) using similar methodology reports the prices on various personal data, including government ID, location, health history, digital communication, and so on. However, all prices were obtained via survey responses. So, the prices found by the study are again "stated" values, not actual values.

The value of individuals' location data was examined in Cvrcek, Kumpost, Matyas, and Danezis (2006) and Danezis, Lewis, and Anderson (2005). In both studies, participants were

provided with deceptive information that they might be paid by allowing their mobile phone location to be monitored for a certain period. They were asked to submit quotes for the payment they would require if their locations were to be monitored. The results show a mean value of £27.4 in Danezis et al. (2005) and €43 in Cvrcek et al. (2006). However, both studies terminated after receiving the quotes. So, the actual exchange of location data and payment did not take place. As a result, the amounts found in the studies are again "stated" values, not the actual values. This is problematic because prior studies on consumer valuation of nonmarket goods have found that people's valuation can be quite different in hypothetical settings than in situations when money is truly on the line (List & Gallet, 2001; Lusk & Shogren, 2007).

Auctions have been recognized in the literature as an effective way to facilitate exchange of personal information and payment (Preibusch, 2013; Dandekar, Fawaz, & Ioannidis, 2014; Ghosh & Roth 2015). An experimental auction to study the monetary value of people's age and weight information was conducted in Huberman, Adar, and Fine (2005). The study involves 127 participants, including students and full-time employees. The participants were paid based on the reverse second-price auctions (where the individual bidding the lowest price wins an auction and is paid the second lowest price). The study finds the average valuation of $57.56 for age data and of $74.06 for weight data. In another study of location data valuation (Staiano et al., 2014), 60 participants were given smartphones that not only can record their location and other usage data but also allow them to enter bid price to sell the data. The reverse second-price auction was also used to determine the winner and the winning price. The experiment went on for six weeks. Overall, the median price for daily location data was about two euros (€2). These studies involved actual exchange of personal data and payment, which confirms that individuals are willing to trade their private information for monetary rewards, indicating that auctions are a practical mechanism to value personal data.

Another study using experimental auction assesses the monetary value of personal information on 168 general Internet users, using again a reverse second-price auction (Carrascal, Riederer, Erramilli, Cherubini, & de Oliveira, 2013). Data items examined include users' basic personal data, such as age, gender and address, as well as their online activity data, including searching, shopping and social-networking data. Different data items were auctioned separately. The data were collected using a plugin application installed in participants' browsers during a 2-week study period. The median value for the basic personal data is found to be €25 ($36), which is higher than that of any other category of online activity data. However, the study focused more on some specific online behavior data than general personal data such as identity and demographic data.

The studies that use auctions to elicit participants' valuations (Huberman et al., 2005; Carrascal et al., 2013; Staiano et al., 2014) all use a single-item second-price auction, which is incentive compatible (or truth telling) in that there is no incentive for a bidder to submit a price that deviates from the bidder's true valuation. However, because each second-price auction has only one winner, it is difficult to attract a large pool of participants to bid in the study. Therefore, these studies were conducted by running multiple second-price auctions sequentially; each auction normally included between 10 and 20 bids. The use of multiple

sequential auctions has some limitations. It is well documented that sequential auction prices tend to follow a declining pattern, making valuation inconsistent (Ashenfelter, 1989; McAfee & Vincent, 1993; Krishna, 2010). In addition, with sequential auctions and repeated bidding activities, participants in the above studies knew that they were in an experiment. Due to the Hawthorne effect (Adair, 1984; McCambridge, Witton, & Elbourne, 2014), the bid price might not reflect well the actual trade-off between privacy and monetary incentive in a real-life setting. To accurately assess the value of personal information, it is desirable to use an auction mechanism that is reliable for estimating the price value in an experiment as close to real-life scenarios as possible.

Another common limitation of existing studies on personal data valuation is that they are all value data items in isolation. For example, Carrascal et al. (2013) assess the individual values of many items, such as age, gender, address, search keywords, and online browsing activities. But they do not consider the combined value of these items. Clearly, the combined value is not a simple sum of the individual values. For example, the combined value of a person's name, email, and phone number is unlikely to be the sum of the three items' individual values because knowing one or two of them should be sufficient to find the other items. On the other hand, if the name is associated with the person's financial or medical information, the combined value will be much higher. This work aims to assess the combined value of different types of personal data items. To the best of our knowledge, this problem has not been studied in literature.

RESEARCH FRAMEWORK

As the primary objective of this research is to develop an economic mechanism for valuing personal information, we first describe our personal information valuation model. We then discuss a set of related variables representing the individuals' self-disclosure behaviors, their stated privacy preferences and attitudes, and their demographic characteristics.

Personal Information Valuation Model

Unlike the behavioral privacy models that typically include latent constructs, economic valuation of personal information typically uses variables that are observable or measureable, as described in the related work in the previous section. The personal information valuation model proposed in this study also consists of variables that are directly observable or measureable. More specifically, our model considers personal data that are generally collected and stored in the databases of various organizations, including the original data owner organizations or data brokers. These data can be distributed as information goods and different attributes of the data have different values and privacy implications. To make our research more generalizable, we use attributes that are not dependent on specific application contexts. These attributes include individuals' name, phone number, email address, home address, age, gender, race, education, occupation, marital status, income, and so on.

In our model, personal information is valued as a combined aggregate of the numerous data attributes of an individual. We use experimental auctions to elicit individuals' valuation of their combined data. To carry out the study, it is necessary to consider what attribute values

an individual needs to provide for participating in the auction and how bid prices are related to the attributes. On the one hand, it may not be appropriate to ask an individual to submit a single bid for all values of the requested attributes, as some of the attributes might not be applicable to the individual or the individual may be unwilling to bid for some attributes. On the other hand, it is also not suitable to ask a person to specify a price for each attribute independently. For example, it is not interesting to assess the value of a person's name without other information or the value of a person's salary information without the person's identity, because these values independently do not have meaningful privacy context.

A type of auction, called combinatorial auctions, appears to be useful to deal with the above situation. Combinatorial auctions allow bidders to place bids on combinations of items, rather than just individual items (Cramton, Shoham, & Steinberg, 2006). Using combinatorial auction, it is possible for a participant to submit a bid for a combination of attributes that the participant is willing to disclose. Although theoretically appealing, sophisticated combinatorial auctions are not suitable for our study for several reasons. First, it is well recognized that the winner determination problem for combinatorial auctions is computationally intractable in general (de Vries & Vohra, 2003). Second, although there exist some special tractable cases of the winner determination problem (Rothkopf, Pekeč, & Harstad, 1998), it is difficult to explain to the participants in an experimental study setting how solution algorithms work. Third, even if a combinatorial auction can be implemented for the study, the data collected may not be useful for analysis due to an exponentially large number of valuation outcomes. For example, if a study includes 15 attributes, then there will be as many as $2^{15} - 1 = 32{,}767$ possible attribute combinations. The number of bids for each combination is likely to be small unless the number of participants is sufficiently large or the number of bids submitted from each participant is sufficiently large. Perhaps, because of these reasons, none of the existing auction studies of personal data valuation implemented combinatorial auctions.

To simplify the valuation scenario, we group the personal data into a few categories. In the data privacy literature, the attributes of personal data are often classified into four categories (Sweeney, 2002; Fung, Wang, Chen, & Yu, 2010; Li & Sarkar, 2011; El Emam, 2013; EU, 2016): (i) *direct identifiers*, such as name, phone number, and social security number (SSN); (ii) *quasi-identifiers*, which can be used to potentially identify an individual, such as date of birth and zip code; (iii) *nonprivate attributes*, which can often be obtained from public sources and are normally not considered sensitive, such as age, gender, race, and other demographic data; and (iv) *private attributes*, which contain sensitive private information, such as salary, disease, mobile location, and online activity data. Based on this classification in the literature, we have designed the auction that asks a bidder to submit a price for each of the following four categories: Identifiers (**I**), Quasi-identifiers (**Q**), Demographics or nonprivate (**D**), and Private (**P**) information. Specific data items included in the four types are given in Table 1.

In the identifier type, we do not include SSN due to legal concerns. In the quasi-identifier type, we include two attributes, date of birth, and five-digit zip code, based on two grounds: (i) The Privacy Rule in HIPAA (Health Insurance Portability and Accountability Act) specifies 18 categories of patient information that must be removed before the health data

are released to a third party (DHHS, 2000). Sixteen of them are direct identifiers such as name and SSN; the other two are date of birth (which must be truncated to year of birth) and five-digit zip code (which must be truncated to keep the first three digits only). (ii) A well-known study (Sweeney, 2002) found that 87% of the population in the United States can be uniquely identified with three attributes—date of birth, five-digit zip code, and gender—which are accessible from voter registration records available to the public. Date of birth and zip code are also interesting in that they can also be used as demographic attributes. As such, we include their HIPAA-compliant format (age and first three digits of zip code) in the demographics category. For the private information category, there can be many attributes, including health and medical information, financial and tax information, online purchase activities, sexual orientation, and so on. However, they are mostly context-dependent and some of them are overly sensitive to collect for this study. To make our research easily generalizable, we avoided context-dependent attributes and selected only income and credit scores, which are common to most consumers.

The four types of attributes above have different usage values and privacy implications. Identifiers are generally used by organizations to contact individuals and to send marketing and promotion materials. Quasi-identifiers can be used as both potential identifying attributes and demographic attributes. Demographic and private attributes are very useful for business analytics applications, as well as for social science and business research. While demographic attributes are generally not considered privacy sensitive and can be obtained with relatively low cost, private attributes are often highly valued by the individuals due to their sensitive nature.

In our valuation model, the value of personal information for an individual, V, is considered a composite function of the values of the four types of attributes. That is,

$$V = f(V_I, V_Q, V_D, V_P), \tag{1}$$

where V_I, V_Q, V_D, and V_P are the value of the identifier, quasi-identifier, demographics, and private attributes, respectively. An important characteristic of this valuation function is that it is hard to independently assess the individual value of each type, V_I, V_Q, V_D, or V_P, without information on at least another type. For example, a person's income and credit score data alone without identity or demographic information do not seem to have much value. We have also explained earlier that identity information alone has little value. To deal with this issue, our auction study requires the participants to submit their personal information and corresponding bid for at least two types. In this way, one type of information is tied to at least another type and related bid prices represent a meaningful valuation of the personal information.

There is not any existing work that studies this functional form for personal information valuation. We thus begin with a simple scenario. In this study, we consider the value of personal information as a linear combination of the values of the four types of attributes. That is,

$$V = k_I V_I + k_Q V_Q + k_D V_D + k_P V_P. \tag{2}$$

As explained, we require function (2) to have at least two V_j ($j = I, Q, D, P$) terms.

The Westin Index for Privacy Preferences and Attitudes

Consumer privacy preference and attitude have been studied extensively in the literature. There are various survey-based instruments for measuring privacy attitude and preference. One of the well-known instruments is the Westin privacy segmentation index (Kumaraguru & Cranor, 2005). This index evaluates individuals' privacy attitudes and preferences based on their answers to three simple questions related to their personal data disclosures and organizations' collection and use of personal data. The three questions succinctly represent three important dimensions of consumers' privacy preferences and attitudes, which are: (i) consumers' ability to control the disclosure of their personal information, (ii) organizations' privacy awareness and practice in handling personal data, and (iii) effectiveness of existing laws and policies in protecting consumer privacy (Kumaraguru & Cranor, 2005). The instrument was developed and used by Dr. Alan Westin and his colleagues for three decades in more than 30 privacy surveys and studies for measuring privacy preferences and attitudes in academic research and practical context (Kumaraguru & Cranor, 2005). The Westin survey has widely been used as a benchmark against which other privacy researchers can compare their own study results (Jensen, Potts, & Jensen, 2005; Kumaraguru & Cranor, 2005; Krasnova, Hildebrand, & Guenther, 2009; Woodruff et al., 2014). In this study, we have chosen the Westin index survey instrument because of its simplicity, longevity, proven validity, extensive use, and not limited to a specific context.

The Westin segmentation survey includes the following three questions/statements:

Q1. Consumers have lost all control over how personal information is collected and used by companies.

Q2. Most businesses handle the personal information they collect about consumers in a proper and confidential way.

Q3. Existing laws and organizational practices provide a reasonable level of protection for consumer privacy today.

In Westin's privacy segmentation survey, the participants were requested to provide an answer to each question from four possible choices: (i) Strongly Disagree, (ii) Somewhat Disagree, (iii) Somewhat Agree, and (iv) Strongly Agree. Based on the participants' responses, they are classified into three segments: privacy fundamentalists (who have the highest privacy concerns), privacy pragmatists, and privacy unconcerned (who have the lowest privacy concerns). These segments can then be used for direct marketing or privacy research purposes.

Many prior studies have shown that consumers with higher privacy preferences and attitudes have higher valuation of their personal data. Krasnova et al. (2009) have clustered the individuals into three categories based on the Westin segmentation survey and other survey questions and measured their valuation of personal data. They found that individuals with stronger privacy preferences and attitudes have higher valuation of their personal data than those with weaker preferences and attitudes. Schreiner and Hess (2015) found

that individuals' WTP for privacy is positively related to their privacy intensions and attitudes. Grossklags and Acquisti (2007) observed that privacy-concerned individuals request higher payments for their information and are willing to expend more money to be protected than privacy-unconcerned individuals. Similarly, Hann, Hui, Lee, and Png (2007) found that "privacy guardians" (fundamentalists) attach higher values to their information than "information sellers" (privacy unconcerned). Related to Q2 of the Westin survey, Spiekermann and Korunovska (2017) found that higher organizational privacy concerns are positively related to individuals' valuation of their personal data. Based on these empirical research findings, we have

> **Stylized Fact 1**: Consumers with more conservative privacy preferences and attitudes have higher valuation of their personal information than consumers with less conservative privacy preferences and attitudes.

We should point out that several privacy studies (Jensen et al., 2005; Premazzi et al., 2010; Woodruff et al., 2014) have found results that are inconsistent with Stylized Fact 1. These studies found that individuals' stated privacy preferences and attitudes do not necessarily align with their privacy valuation, a phenomenon called *privacy paradox*. This issue will be further discussed later in the article.

Self-Disclosure and Privacy Control Behaviors

In survey research and practice, it is well known that participants with high privacy concerns tend to refuse to respond to questions involving personal information (Korkeila et al., 2001; Shoemaker, Eichholz, & Skewes, 2002; Joinson, Paine, Buchanan, & Reips, 2008). This leads us to study the relationship between consumers' data hiding/disclosing behavior and economic valuation. Spiekermann and Korunovska (2017) found that less privacy controls over personal data generally lead to higher economic valuation of the data and suggest that consumers with high valuation of their data may withdraw from the data market by not providing their data for valuation. Premazzi et al. (2010) found that consumers' disclosure behaviors tend to be conservative when monetary incentives are insufficient. Goldfarb and Tucker (2012) used consumers' refusals to provide personal information in surveys to measure consumers' privacy values for their study of shifts in privacy concerns. Likewise, a participant with high valuation of their data in our study can also choose not to submit a bid for selling certain types of their data. Thus:

> **Stylized Fact 2**: Consumers who choose not to provide their personal information have higher valuation of their information than those who provide the information.

Privacy researchers have observed that consumers are willing to pay for better self-disclosure controls. Schreiner, Hess, and Faranak (2013) found that consumers are willing to pay for a premium version of Facebook or Google that has better privacy controls if such a premium version is offered. The price they are willing to pay depends on the perceived privacy control functionalities. Krasnova et al. (2009) found that privacy-concerned social network users are willing to pay more for privacy controls than privacy-unconcerned users, and better self-disclosure control features have higher valuations. The self-disclosure and privacy control behaviors can be measured based on individuals' responses to the related questions (John, Acquisti, & Loewenstein, 2011; Woodruff et al., 2014). Krasnova et al.

(2009) and Hallam and Zanella (2017) have developed survey instruments for studying self-disclosure and privacy control behaviors. Their privacy surveys include privacy behavior scenarios and action-based questions regarding how users self-disclose their personal information with their friends and online social community. We have therefore adapted our self-disclosure behavior questions from their questionnaires to form our next three questions.

Q4. (ShareInfo) How often do you share your salary (or grade for student) information with your friends?

1. Always 2. Often 3. Don't remember 4. Not often 5. Never

Q5. (Facebook) On Facebook, I share my posts, photos, or email address with

1. Public 2. Friends-of-friends 3. I don't use Facebook 4. Friends 5. Nobody

Q6. (Twitter) On Twitter, my privacy setting allows

1. Show location 2. Let others find me 3. I don't use Twitter

4. Approved followers 5. Protect my Tweets

In the above answers, one represents the least conservative self-disclosure behavior, while five represents the most conservative behavior. The responses to these questions are used to examine the following relationship:

> **Stylized Fact 3**: Consumers with more conservative self-disclosure behaviors have higher valuation of their personal information than those with less conservative self-disclosure behaviors.

Observations about Consumer Demographics and Privacy

Prior studies have investigated the relationships between consumer privacy and some demographic attributes. In general, studies have consistently found that older consumers tend to be more concerned about privacy than younger consumers (Graeff & Harmon, 2002; Janda & Fair, 2004; Varian, Wallenberg, & Woroch, 2005; Zukowski & Brown, 2007; Laric, Pitta, & Katsanis, 2009; Joinson, Reips, Buchanan, & Schofield, 2010; Goldfarb & Tucker, 2012), and women tend to be more privacy-sensitive than men (O'Neil, 2001; Graeff & Harmon, 2002; Janda & Fair, 2004; Laric et al., 2009; Hoy & Milne, 2010; Joinson et al., 2010). Some studies have also examined the relationships between consumer privacy and income, education, and race, but the results are inconsistent or inconclusive (Li, 2011).

The above relationships between privacy and age or gender were found in a variety of privacy contexts, including direct marketing, social networking, online shopping, health data sharing, mobile activity, and so on. The observations and findings on the relationships are consistent and well accepted in the privacy literature (Li, 2011). In this study, we examine whether the economic valuation of personal data is consistent with these observations.

> **Stylized Fact 4**: Older consumers tend to have higher valuation of their personal information than younger consumers.

> **Stylized Fact 5**: Female consumers tend to have higher valuation of their personal information than male consumers.

We use “tend to” in the above two statements because while the above relationships between privacy and age or gender have been observed in the majority of the studies, a small number of studies have nevertheless found the opposite relationships. *We should point out that the purpose of Stylized Facts 1 through 5 is to help to examine if the results of our economic valuation approach are consistent with the empirical findings from prior related studies, rather than to build new theories*.

METHODOLOGY AND DATA COLLECTION

As discussed in the related work section, an auction is an appropriate and practical method for valuing personal information. However, existing auction studies typically use multiple sequential auctions, which may not measure well individuals’ actual value of their personal data in real life scenarios. Therefore, we consider using a single auction mechanism in this study. To attract a sufficient number of participants, this single auction should be able to result in multiple winners. One of the auction mechanisms with this feature is the classic Vickrey–Clarke–Groves (VCG) auction. The VCG auction is a theoretically ideal mechanism because it is incentive compatible. However, the VCG mechanism is hard for participants to understand. There are also other various reasons that make it impractical to implement the VCG auction in a valuation study (Rothkopf, 2007). Another well-known incentive-compatible mechanism is the Becker–DeGroot–Marschak (BDM) auction, where the winning price depends on a randomly drawn number. Under the budget constraints, however, it is difficult to determine the distribution to be used for the random price generator. It is also difficult to explain to participants why a certain distribution is used. BDM auctions have been used to value Facebook user data in Bauer, Korunovska, and Spiekermann (2012) and Schreiner and Hess (2015). However, both studies terminated after receiving the bids. So, the bid prices were hypothetical and the values found in the studies are “stated” values. To avoid the limitations of the VCG and BDM auctions, we propose using GSP auction mechanism (Edelman, Ostrovsky, & Schwarz, 2007; Varian, 2007), which is a single auction mechanism that is relatively easy to implement.

The GSP mechanism is widely used in sponsored search auctions (Edelman et al., 2007; Lahaie, Pennock, Saberi, & Vohra, 2007). It has been used by Google and some other search engines to draw keyword advertisements online and to determine the payment prices of advertisers when their advertising links get clicked (Lahaie et al., 2007; Ghose & Yang, 2009). GSP is not an exactly incentive-compatible mechanism (Edelman et al., 2007; Varian, 2007). However, it has been shown that GSP is approximately an incentive-compatible mechanism in large sample cases (Liu & Li, 2016). Practically, GSP is very easy to implement. It can be used to award a large number of bidders at one go. In our problem context, the bidders participate in an auction to sell rather than to buy. So, the original GSP auction is reversed, as well as the method of winner determination. In a reverse GSP auction for purchasing data, the bidders submit bids that represent the minimum payment amount they can accept for selling their data. The winners are selected by the bids from the lowest to the highest. The person with the lowest bid is paid an amount equal to the second lowest bid, the person with the second lowest bid is paid an amount equal to the third lowest bid, and so on. This continues until the budget of the auctioneer/buyer for acquiring data runs out. In this way, all data can be acquired in a single auction.

We sent a recruiting letter by mail and email to individuals whose contact information was collected from various sources, including the voter registration lists of several states in the United States, a consumer email list from a data aggregation company, and several email lists from a university. We then developed a Web site for the study and provided an access to the site for potential participants. An important message on the study Web site was that we were contacted by several marketing companies and data vendors to help them to identify consumers who were willing to sell their personal data to the companies for legitimate business use. This information is deceptive because we had not been contacted by any third party. This deceptive scenario, however, is necessary because the objective of the study is to obtain the participants' valuation of their personal data that will be truthfully reported under such a scenario. Because the deception was involved, the study plan was subject to a full review by the Institutional Review Board (IRB) and eventually received the IRB approval. The study was conducted in two phases: phase 1 with student participants and phase 2 with nonstudent participants. As students tend to bid with lower prices than nonstudents, this two-phase procedure allows nonstudents to have a better chance to win than a single-phase experiment. The participants did not know there were two phases; they also did not know the information about the other participants.

The study procedure was designed as follows. First, the participants were presented with the instructions for the study, which explains the background of the study (i.e., several data collectors were interested in purchasing data collected in this study), what specific information is requested in the study, and how the participants are paid (which is explained using an illustrative example of a reverse GSP auction). Then, the participants were required to complete a personal information form, which includes the data items listed in Table 1. To reduce the chance of participants providing false information, they were required to attest that all the information provided is true and correct; they were also told that the information they provide in this study will be verified to win the auction. The participants then entered the bid price, one price for each of the four categories, as the requested compensation for providing their personal data (they were asked to complete and bid for at least two categories).

The participants were told that the total budget amount that the data collectors would spend for each category is $500, and the participants were paid based on their bids in each category from the lowest to the highest until the budget was used up. The $500 budget for each category was set based on the results of prior studies (Hann et al., 2002, 2007; Danezis et al., 2005; Huberman et al., 2005; Cvrcek et al., 2006; Carrascal et al., 2013; Staiano et al., 2014). As described in the related work section, the valuations of personal data from the participants in these prior studies were all significantly lower than $500.

As part of this study, we also mentioned that the data collectors are interested in purchasing the following information items about the participants: (i) SSN, (ii) total federal income tax paid last year, (iii) total medical claims last year, and (iv) body weight and height. This is an optional part and the participants were asked to specify a price for each item. Because these data are very sensitive, we did not intend to collect them but would like to get some idea about how consumers would value these data. We clearly instructed the participants not

to enter the actual information (i.e., SSN, income tax, etc.) and emphasized that no payment would be involved in this part.

This study also intends to examine the relationships between individuals' valuation of personal data and individuals' stated privacy preferences and attitudes and their self-disclosure behaviors. Therefore, the participants were required to provide answers to survey questions Q1 through Q6, in addition to providing personal data and price information. To provide sufficient time for participants to consider the situation, the study Web site was open 10 days for submission. Overall, 225 people participated in the auction study, but data entered by a few participants were invalid. As a result, there were 218 records useful for the study. Among them, 88 participants received payments based on their bids.

DATA ANALYSIS AND RESULTS

We first provide descriptive and summary statistics. We then present the results of correlation and multiple regression analysis, followed by the results of partial least squares (PLS) regression.

Descriptive and Summary Statistics

Summary statistics about the participants are given in Table 2. In general, the demographic distributions of the participants (sample) compare fairly well to those of the general public (population). A few statistics appear to have some bias, such as relatively high percentages of Asian and Never Married, and low average Age, which are likely caused by a sizable proportion (36.2%) of the student participants. Nevertheless, the participants appear to reasonably represent the general public. It can be seen that the race distribution is highly unbalanced, and thus, its values are grouped into two categories (white and others) in later analysis. Similarly, marital status values are also grouped into two categories (never married and others) in later analysis.

Summary statistics for some attributes are difficult to report in the table format and thus are described in words here. In terms of location, the participants were located in 22 states in the United States; one participant was in Canada. The occupation data were entered by participants instead of precategorized. It thus includes a wide range of occupations, including managers, engineers, teachers, police officers, health professionals, human resource staff, IT professionals, as well as students. We attempted to group the occupation data into a few categories in different ways but they were still not very useful for analysis. Therefore, the occupation data were not used for further analysis. In retrospect, the data could be more useful if participants were restricted to selecting from limited predefined categories of occupation. There are many missing values for credit score (a participant was allowed to not enter a credit score if an explanation was provided). As a result, the credit score data were used for descriptive analysis only.

Table 3 provides average bid prices for each of the four data types, including those grouped by gender, race, marital status, and student status. It is observed that, in general, the average bid prices for all four data types are lower for males (than females), nonwhites (than whites), never married (than others), and students (than nonstudents). A multivariate analysis of

variance (MANOVA) was performed for each of these cases. The results indicate that the differences are significant for the student status case (at $\alpha = .01$). They are not significant for the other three cases in the MANOVA test (at $\alpha = .1$). We also performed two-sample *t*-tests, as well as their nonparametric counterpart Mann–Whitney tests, on each price type considering gender, race, marital status, and student status, respectively. The results of these tests were similar to those of MANOVA, although price differences in gender were somewhat more significant for some types in Mann–Whitney test. It is clear that among the four types of information, the private information (which includes income and credit score) has the highest valuation on average.

More than 80% of the participants specified prices for the four optional items, i.e., prices for SSN, total income tax paid last year, total medical claims last year, and body weight and height. These prices exhibit very large variations, however. The median prices are $350 for SSN, $100 for tax data, $75 for medical data, and $30 for weight and height. As these items were not subject to auction, they are not included in further analysis. However, the information may be useful for future studies that focus on the related context, e.g., medical or financial data privacy.

As mentioned earlier, given that some participants may not want to bid to sell certain types of data, we allowed the participants to enter and bid for a minimum of two data types. It is interesting to examine which types of data consumers are less willing to bid. Table 4 shows the percentage of incomplete data by the participants. The type of data having the highest percentage of incomplete values is the private data (16.06%), which includes income and credit scores, followed by identifier data (12.84%). This suggests that the participants perceive these two types of information to be most sensitive and/or most difficult to assign an economic value. Indeed, revealing these two types of information leads to direct disclosure of a participant's private information with identity, which appears to be more sensitive than disclosure of any other combination of two data types. To compare the bid price with the possible price from those who did not bid, we required that a participant who did not bid for certain types of data enters a price that the participant would have specified if he or she had participated in the bidding. We call this price "nonbidding" price (because the participant did not provide relevant data to bid). The last two columns of Table 4 show the bidding price and nonbidding price for each type of data. It is clear that nonbidding prices are substantially higher than the bidding price for each type. MANOVA and two-sample *t*-tests confirm that all differences are significant at $\alpha = .05$. This suggests that consumers who avoid bidding tend to have higher valuation of their related data than those who bid.

Correlation and Multiple Regression Analysis

Table 5 provides the correlations of all variables effectively considered in the valuation model. It is observed that the four types of prices are highly correlated. The correlations between Income and the four prices are very low and are in mixed directions. Prior studies on the relationship between income and privacy concerns have provided mixed results: some found that consumers with higher income tend to have a higher level of privacy concerns (Graeff & Harmon, 2002; Varian et al., 2005), while others found the opposite

result (O'Neil, 2001; Zukowski & Brown, 2007). Our result also demonstrates that there is no clear relationship between income and valuation price. We have thus removed Income from further consideration. Note that after grouping, Race (with White = 0, Others = 1) and Marital Status (with Never Married = 0, Others = 1) are binary variables. We include them and Gender (Male = 0, Female = 1) in the correlation analysis mainly to examine their directions of correlations with prices.

We now consider estimation of the valuation model. As discussed earlier, it is not appropriate to simply add the prices of the four data types to form a "total price" for estimation (in fact, some participants did not bid for all four types). Thus, we first consider estimating each of the four price values individually, using ordinary least squares (OLS) regression. A separate OLS regression was formed, each using one of the four prices as the dependent variable. Independent variables for each regression include Age, Gender, Race, Marital Status, and Education. It follows from the previous discussion that whether or not participants provided information for a given data type considerably affects the estimation of the related price. Thus, the independent variables also include a binary variable, Complete (where Complete = 1 indicates that the participant provided completed data of a given type). The distributions of the original prices are not normal. So, a square root transformation was applied to the prices (we also performed logarithm transformation but it is not as effective as square root transformation in normalizing the values). The pairwise relationships between transformed prices and the numeric independent variables are approximately linear. Consequently, the OLS regression models are:

$$\begin{aligned}\text{SqrtPrice}_j = {} & \beta_0 + \beta_1\ \text{Age} + \beta_2\ \text{Gender} + \beta_3\ \text{Race} \\ & + \beta_4\ \text{Marital} + \beta_5\ \text{Edu} + \beta_6\ \text{Complete}_j,\end{aligned} \tag{3}$$

where $j = I, Q, D,$ and $P,$ respectively. For comparison purposes, we have also performed the OLS regression using the original price values. The regression results based on model (3) are reported in Table 6. Related discussion is given later, together with the results from the next model.

Next, we examine the relationships between price values and all variables of interest, including privacy attitude variables (i.e., Fundamentalist, Pragmatist, and Unconcerned, related to Q1 through Q3) and self-disclosure behavior variables (i.e., ShareInfo, Facebook, and Twitter, related to Q4 through Q6). The OLS regression models are:

$$\begin{aligned}\text{SqrtPrice}_i = {} & \beta_0 + \beta_1\ \text{Age} + \beta_2\ \text{Gender} + \beta_3\ \text{Race} + \beta_4\ \text{Marital} + \beta_5\ \text{Edu} \\ & + \beta_6\ \text{Complete}_j + \beta_7\ \text{Fundamentalist} + \beta_8\ \text{Pragmatist} \\ & + \beta_9\ \text{ShareInfo} + \beta_{10}\ \text{Facebook} + \beta_{11}\ \text{Twitter}\end{aligned} \tag{4}$$

where Fundamentalist and Pragmatist are binary variables (i.e., Fundamentalist = 1 for a Fundamentalist, Pragmatist = 1 for a Pragmatist, and if both values are zero, the person is an Unconcerned). The Cronbach's alpha for the questionnaire items is .74, which suggests an acceptable reliability.

The regression results based on model (4) are reported in Table 7. Some important observations can be made based on the results in Tables 6 and 7. First, we observe that

the positive relationships between Age and the prices of the quasi-identifier, demographic, and private data types are statistically very significant, which is consistent with Stylized Fact 4 that older consumers value their personal information higher than younger consumers. Second, the positive relationships between gender and those prices are statistically significant for most cases, which is consistent with Stylized Fact 5 that females value their information higher than males. Race is significantly related to only one price type—white consumers tend to value their demographics data (which include their race) somewhat lower than the others. Also, education is somewhat significantly related to only one price (for identifier type data). Both results may explain why prior studies on these aspects are inconclusive. Marital Status is not significantly related to any price. We observe that the *R*-squared values for all cases are fairly low, suggesting that a large portion of price variations are not explained by the variables considered. However, this does not affect our interpretation above regarding the relationships between the prices and independent variables.

It is interesting to observe from the coefficients of the Complete variable that the relationships between individuals' data hiding behavior and most of the prices are statistically very significant. The negative coefficients of this variable indicate that participants who submitted their personal data tend to have a lower price than those who did not. This result, together with the price comparison results shown in Table 4, are consistent with Stylized Fact 2 that consumers who hide their personal information have higher valuation of their personal data than those who disclose. Note that the coefficient of this variable for the price of demographic data cannot be estimated because there is no data for the independent variables (e.g., Age, Gender, etc., which are all demographic) for those who chose not to bid for this type. It is also observed from Table 7 that the relationships between most of the prices and individuals' self-disclosure behaviors in Facebook and Twitter are statistically significant. The relationships between prices and ShareInfo do not look statistically significant. We believe that this is caused by the ambiguity in the words designed to answers Q4 ("often" and "not often" is not defined and thus can be interpreted differently by different participants). In general, there are strong evidences (incomplete data, and Facebook and Twitter settings) to support that consumers' valuation of personal data is consistent with their self-disclosure behaviors (Stylized Facts 2–3).

On the other hand, Table 7 shows that there is no statistically significant relationship between stated privacy preferences and any price. The negative coefficients associated with the Pragmatist variable implies that Privacy Pragmatists value their personal information lower than the Privacy Unconcerned, which is opposite to what Stylized Fact 1 states. Therefore, stated privacy preferences are not consistent with the economic valuation of personal data in this study. This is not particularly surprising because, as we mentioned earlier, some empirical studies have found results that are inconsistent with Stylized Fact 1.

Partial Least Squares Regression for Valuing Personal Information

While OLS regression can estimate the components of the valuation function, V_I, V_Q, V_D, and V_P in Equation (2), it cannot estimate the composite valuation function V. To develop a valuation model for measuring the composite value, we turn to *PLS* regression. PLS is

a regression technique that can handle multiple dependent variables (Wold, Sjöström, & Eriksson, 2001; Hastie, Tibshirani, & Friedman, 2009).[1] It is particularly effective when the dependent and/or independent variables are highly correlated. This capability is achieved by extracting from the data a few latent components that have the best predictive power. It can be seen from Table 5 that the correlation between each pair of the dependent variables (prices) is very high. So, PLS is an appropriate method to estimate the composite valuation function.

Let n be the number of observations, p be the number of independent variables, and **X** be an $n \times p$ data matrix for the independent variables (which is the same as the **X** in OLS). Let q be the number of dependent variables and **Y** be an $n \times q$ matrix for the dependent variables (while the OLS counterpart **y** is $n \times 1$). The PLS model can be essentially written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{5}$$

where **B** is a $p \times q$ coefficient matrix (while the OLS counterpart $\boldsymbol{\beta}$ is $p \times 1$) and **E** represents errors. In OLS, solution for the regression coefficients $\boldsymbol{\beta}$ can be written as a closed-form expression. In PLS model (5), however, **B** is estimated iteratively by considering some latent components **T** (similar to factors or principal components in multivariate statistics). More specifically, **B** can be decomposed as **B** = **WC**, where **W** and **C** represent the weights of **X** and **Y** on the latent components **T**, respectively. The values of **W** and **C** and thus **B**, as well as latent component scores, can be derived iteratively one component at a time. In our problem, we consider only one latent component that generalizes the four types of prices. This component, which can be called the composite price, is expressed as a weighted sum of the four prices as in Equation (2), and the weights can be found from the PLS solution.

Applying the PLS procedure for the data, we obtain the following composite valuation function:

$$V = 0.497V_I + 0.405V_Q + 0.447V_D + 0.624V_P. \tag{6}$$

It is clear that the combined value of personal data is not a simple sum of the value of each data type, either in the scale of the total amount or in the weights of different data types. Model (6) represents the combined valuation with all four types of data. To build a valuation model involving fewer types of data, we can run the PLS procedure using only the price data corresponding to the relevant types. Model (6) suggests that private data have the highest weight in the composite valuation function, followed by identity data. Demographic and quasi-identifier data have relatively lower weights. This result is rather (though not completely) consistent with the results in Table 4 in terms of relative importance of the four types of data. It is also consistent with the results of some prior studies (Phelps et al., 2000).

[1]PLS regression used here is different from PLS path modeling. The former is a prediction technique, while the latter is a research method for testing research hypotheses. There are some concerns about using PLS path modeling as a research method, which are irrelevant to the use of PLS regression for prediction and estimation in our study.

Based on the estimated model (6) and average price value for V_I, V_Q, V_D, and V_P given in Table 3, the overall average price value for the participants who provided all four types of personal data is:

$$V = 0.497(\$89.76) + 0.405(\$88.94) + 0.447(\$78.91) + 0.624(\$117.78) = \$189.40\,. \tag{7}$$

A recent study by IBM and Ponemon Institute (2017) reported that the average cost incurred for each lost or stolen record containing sensitive information was $225 in the United States. Our estimated value ($189.40) is quite comparable to this amount, as well as to the amount specified in the bill introduced by senators Warren and Warner (2018).

Performance Evaluation

To evaluate the effectiveness of the proposed GSP auction mechanism, we conducted a performance evaluation study on the collected data. Based on our literature review, prior studies have used the reverse second-price auction (Huberman et al., 2005; Carrascal et al., 2013; Staiano et al., 2014) and the BDM auction (Bauer et al., 2012; Schreiner & Hess, 2015) for valuing personal data. As the BDM auction can result in multiple winners in a single auction while the second-price auction cannot, we used the BDM auction as a baseline for comparison. Because both BDM and GSP auctions are approximately incentive compatible truth-telling mechanisms, a participant is expected to bid with the same price under either auction.

As mentioned earlier, our data set includes 218 participants, of which 88 participants received payments based on their bids in the GSP auction. We were authorized by all 218 participants to use their data for this research even though many of them did not receive payment. If the GSP auction were carried out by a data vendor/purchaser for commercial purposes, the vendor with the same budget could only obtain the subset of the 88 records (the unpaid individuals would not provide their data free for commercial use). An effective way to compare the performances of different auction mechanisms is to examine how well the subset of the data acquired by each auction with the same budget represents the entire *reference* data set of the 218 participants (i.e., the "population" who participated in the study). Thus, we applied a BDM auction with the same budget to the reference data set, which selected a subset of 62 records. The subset obtained by the BDM auction was then compared with that obtained by the GSP auction, based on the performance measures defined next.

The first performance criterion is the closeness between the summary statistics of the selected subsets and that of the reference data. We have provided summary statistics for the reference data in Table 2, including the frequency distributions of the categorical attributes Gender, Race, and Marital Status, and the mean values of the numeric attributes Age, Education, and Income. For a categorical attribute with C categories, the closeness is measured by the average difference in frequency (ADF), defined by

$$\text{ADF} = \frac{1}{C}\sum_{k=1}^{C}|f_k - F_k|,$$

where f_k and F_k are the relative frequency (in percentage) for the kth category in the subset and reference set, respectively. For a numeric attribute, the closeness is measured by the average difference in mean (ADM), defined by

$$\text{ADM} = \left|\bar{x} - \bar{X}\right| / \bar{X},$$

where $\bar{x}$ and $\bar{X}$ are the mean of the numeric attribute in the subset and reference set, respectively. Clearly, a small ADF or ADM value indicates a small deviation in the summary statistic from the reference data, which is desirable.

The second performance criterion is prediction error with OLS regression. We divided the reference data into two subsets: the records obtained by an auction served as the training set to build the regression model, and the remaining records served as the test set. The prediction error is measured by the root mean squared errors (RMSE) as follows:

$$\text{RMSE}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{ij} - \hat{y}_{ij}\right)^2},$$

where n is the number of records in the test set, y_{ij} is the bid price from the ith record in the test set for data type j ($j = I, Q, D, P$), and $\hat{y}_{ij}$ is the predicted value for y_{ij}, calculated using the regression model (3) built on the training data. The average RMSE values over the four data types are then calculated and reported.

The results are shown in Table 8. It can be seen that the frequency distributions of Age, Race, and Marital Status from the data collected with GSP are closer to those distributions from the reference data than those collected with BDM. Similarly, the means of Age and Education from the data collected with GSP are also closer to those from the reference data than those collected with BDM. In addition, the prediction error with GSP is smaller than that with BDM. It is clear that the proposed GSP auction mechanism outperforms the BDM auction mechanism in all measures, suggesting that the subset obtained by the GSP auction better represents the reference data than that by the BDM auction. Note that the BDM auction pays the same amount (the winning price) to those whose bids are below the winning price, while the GSP auction pays each participant an amount approximately equal to the participant's bid price. Thus, with the same budget, GSP can acquire more records than BDM, which may explain why the GSP subset represents the reference data better than the BDM subset.

DISCUSSION

Digital economy and business analytics are increasingly relying on the use of consumer data to derive business insights. However, consumers today are increasingly reluctant to share

their personal data due to their privacy dispositions. The Facebook-Cambridge Analytica data scandal has created consumers' backlash to misuse of their personal data without consent (Granville, 2018). To protect consumer privacy, governments in the United States (DHHS, 2000; FTC, 2014) and European Union (EU, 2016) have established regulations and rules, such as HIPAA and General Data Protection Regulation (GDPR), which requires a consent from the consumers and more transparency in the data acquisition and sharing process. The tension between the capability and need for collecting consumer data and legal and ethical responsibilities for protecting data privacy has presented a great challenge for the digital economy companies. A plausible approach to ease this tension is to provide financial reward for consumers to share their data. If consumers are able to exchange their data for a monetary value, they will be more aware of what they are providing consent for secondary use. This study explores and validates the feasibility of this approach.

A key challenge in the consumer privacy research and practice is measuring the value of personal data. This study proposes using the GSP auction to elicit and measure consumers' economic valuation of personal data and in the process to obtain consent from them to use the data. We find that consumers' valuation of their personal data is consistent with their self-disclosure behaviors. The valuation approach also produces results that are consistent with some well-accepted observations about consumer demographics and privacy. The findings suggest that it is theoretically appealing and practically viable to measure the value of personal information using an auction approach.

This study also finds that consumers' stated privacy preferences and attitudes are not consistent with their valuation of personal data. Prior studies have found that people's assessment and valuation are less reliable in hypothetical and inconsequential settings than in situations when money is truly on the line (List & Gallet, 2001; Lusk & Shogren, 2007). This suggests that economic valuation is likely to serve better for measuring information privacy than stated privacy preferences and attitudes. Consequently, this study can have a promising impact on some important privacy research areas such as privacy calculus and privacy paradox, where assessing the value of personal information is vital. By using economic value of personal information for measuring the benefit component in the privacy risk/benefit analysis, it is reasonable to expect some more reliable privacy calculus models to emerge and new insights to be revealed. For example, the willingness to provide personal information construct in the privacy calculus model in Dinev and Hart (2006) could be substituted with the actual economic valuation of personal information.

The proposed personal information valuation model provides firms with a way to estimate the values or costs of consumer data with privacy consideration. In today's data-rich environment, firms usually own a large amount of customer data or can purchase prospects' data from data vendors at very low cost. Consequently, there tends to be a misconception that customer data are so cheap that their acquisition cost is not really a noteworthy factor in making business decisions. However, customers generally value the worth of their personal data much higher than firms due to their privacy concerns. Data breach and privacy violation can incur substantial financial loss and have a negative impact on a firm's bottom-line performance (Ponemon Institute, 2017). Facebook stock lost $80 billion or 18% in market value within 10 days after its data scandal was reported (La Monica, 2018, March 23).

Therefore, organizations collecting and storing consumer data should consider the cost associated with protecting the security and privacy of customer data in assessing the value of customer data.

This study also offers valuable insights for policy decision makers to assess the values of personal data and to charge a justifiable amount of penalty for a data breach and privacy violation. The average value of personal data found in this study is comparable to the amount of penalty specified in the bill introduced by senators Warren and Warner (2018). This is understandable because most of the data attributes acquired in this study (shown in Table 1) are also collected and reported by a credit-reporting agency. This study can be valuable for the government agencies like FTC to assess the financial cost and damage incurred to the consumers affected in a data breach, such as those in the Facebook-Cambridge Analytica case (Sydell, 2018). The proposed economic valuation method can be used in this regard by using data attributes relevant to the Facebook context.

Prior research on personal data access has shown that firms relinquishing personal data sharing control to consumers increase the consumers' trust and their willingness to share data (Bélanger & Crossler, 2011; Smith et al., 2011). Our research supports this finding by showing that consumers are willing to share their personal information when offered a fair value. By selling their data, consumers grant consent to firms to use their data for marketing and personalization. This has important implications under the tightened new laws and regulations for protecting consumer data. For example, the new GDPR regulation has increased data protection for individuals from all European Union countries, giving extensive control of data back to consumers. It requires explicit consent from consumers for data collection and sharing, and regulates how the data can be used by firms (EU, 2016). Companies that acquire data through direct purchase from consumers will garner more consumer trust and brand loyalty over others that do not. Our study provides a viable approach to acquire consumer data with their consent through economic exchanges. Consumers in our study were willing to provide quasi-identifier and demographic data at relatively low costs, which are very useful for business analytics. Therefore, instead of collecting consumer data from secondary sources without the consumers' consent, practitioners can get the data directly through personal data acquisition where consumers willingly share their data, with consent, in return for a value in an economic exchange. We foresee that in the future, personal data exchanges could be formed with an auction-type platform to purchase individuals' personal data.

Our findings provide privacy researchers and practitioners with insights into the effect of data hiding, i.e., consumers who choose not to provide their information have higher valuation of their personal information than those who provide the information. Our study not only confirms the previous findings qualitatively, but also quantifies the amount of difference in monetary terms. As observed from Table 4, the valuations from consumers who chose not to bid are, on average, about twice as much as those from consumers who placed a bid. Importantly, our finding provides an interesting idea for data surveyors and researchers to measure privacy concerns more accurately. Instead of asking consumers to rate their privacy concerns with a Likert-scaled survey, it may be more effective to ask a number of sensitive questions, each with different degrees of sensitivity and invasiveness,

and to allow nonresponse to the questions. A good survey design along this line can be found in John et al. (2011). Such a survey design can encourage truthful response and reduce the problem of privacy paradox where consumers' stated privacy concerns do not match up to their actual behavior in disclosing private information.

A word of caution is in order regarding the overall average estimated value of the participants in Equation (7). It should be noted that the estimated amount cannot be directly compared to various price values of personal data in prior studies and in practice because the elements of the data and application contexts are different. For example, we mentioned that Datacoup paid its customers $8 a month for their personal data, but it is not clear what exact information was acquired and how many months the customers were paid by Datacoup.

CONCLUSIONS AND FUTURE RESEARCH

Our study looks into a quantitative valuation of personal data from the supply side, i.e., from the perspective of individuals releasing personal information in an economic exchange. There is a value gap between individuals' valuations of their personal data and the data brokers' valuations for secondary data. The value gap between what users wish to protect and what social media sites offer was examined recently using a design science approach (Syed et al., 2019). Our investigation on how personal data is valued economically from the perspective of data users instead of the data sellers or brokers could provide valuable insights on consumer data valuation research in the regulatory environments (EU, 2016).

We assess the value of personal data in a broad context, in order to make our research easily generalizable. Therefore, we have used the Westin index to measure consumer's stated privacy preferences and attitudes because it is not limited to a specific context. When the problem context is specifically given, appropriate instruments specific to the context should be more effective for measuring the stated privacy concern. For example, in studying the Internet users' privacy concerns, the IUIPC measure (Malhotra et al., 2004) has been widely used. Future research may focus on some context-dependent privacy problems using more specific measures.

One limitation with our results is that the *R*-squared values are low, which suggests that a large portion of price variations are not explained by the independent variables. To some extent, this is not surprising because the variables considered in the regression analysis (e.g., Age, Gender, Race, Marital Status, and Education) may only be able to explain a small portion of the value of personal information. Future research should consider more, perhaps context-dependent, explanatory variables. It is also worthwhile considering the use of nonlinear models for the valuation of personal information.

ACKNOWLEDGMENTS

The authors are grateful to the department editor, associate editor, and the two anonymous reviewers for their insightful comments and suggestions that have improved the article considerably. This research was supported in part by the National Library of Medicine of the National Institutes of Health under Grant Number R01LM010942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Biography

Xiao-Bai Li is a Professor of information systems in the Department of Operations and Information Systems at the University of Massachusetts, Lowell, USA. He received his PhD in management science from the University of South Carolina. His research focuses on data science, business analytics, data privacy, and information economics. He has received funding for his research from National Institutes of Health (NIH) and National Science Foundation (NSF), USA. His work has appeared in *Information Systems Research*, *Management Science*, *MIS Quarterly*, *Operations Research*, *Journal of the Association for Information Systems*, *INFORMS Journal on Computing*, *IEEE Transactions (TKDE, TSMC, TAC)*, *Decision Support Systems*, *Communications of the ACM*, and *European Journal of Operational Research*, among others. He is serving as an Associate Editor for *Information Systems Research*, *Decision Support Systems*, *ACM Journal of Data and Information Quality*, and *Information Technology and Management*.

Xiaoping Liu is a Visiting Assistant Professor in the D'Amore-McKim School of Business at Northeastern University. He has published in *INFORMS Journal on Computing*, *ACM Transactions on Management Information Systems*, and *ACM Journal of Data and Information Quality*.

Luvai Motiwalla is a Professor of management information system at the University of Massachusetts Lowell. Dr. Motiwalla earned his PhD in MIS in from the University of Arizona. His current research focuses on information privacy, system usage, mobile systems adoption, organizational assimilation of enterprise systems, and behavioral analytics. He has published two books and numerous articles in various information systems journals, and presented frequently at national and international conferences. He has also received grants from NIH, NSF, U.S. DoE, and funding from private foundations.

REFERENCES

Adair JG (1984). The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology, 69(2), 334–345.

Ashenfelter O (1989). How auctions work for wine and art. Journal of Economic Perspectives, 3(3), 23–36.

Bauer C, Korunovska J, & Spiekermann S (2012). On the value of information – What Facebook users are willing to pay. Proceedings of the 2012 European Conference on Information Systems. Atlanta, GA: Association for Information Systems.

Belanger F, & Crossler RE (2011). Privacy in the digital age: A review of information privacy research in information systems. MIS Quarterly, 35(4), 1017–1041.

Carrascal JP, Riederer C, Erramilli V, Cherubini M, & de Oliveira R (2013). Your browsing behavior for a Big Mac: Economics of personal information online. Proceedings of the 22nd International Conference on World Wide Web (pp. 189–200). New York: ACM Press.

Cramton P, Shoham Y, & Steinberg R (Eds.) (2006). Combinatorial auctions. Cambridge, MA: MIT Press.

Culnan MJ, & Bies RJ (2003). Consumer privacy: Balancing economic and justice considerations. Journal of Social Issues, 59(2), 323–342.

Cvrcek D, Kumpost M, Matyas V, & Danezis G (2006). A study on the value of location privacy. Proceedings of the Workshop on Privacy in the Electronic Society (pp. 109–118). New York: ACM Press.

Dandekar P, Fawaz N, & Ioannidis S (2014). Privacy auctions for recommender systems. ACM Transactions on Economics and Computation, 2(3), Article 12 (22 p.).

Danezis G, Lewis S, & Anderson R (2005). How much is location privacy worth? Proceedings of the Workshop on the Economics of Information Security, Cambridge, MA.

Datacoup. (2019). Unlock the value of your personal data. Accessed October 5, 2019, available at http://datacoup.com

Department of Health and Human Services (DHHS). (2000). Standards for privacy of individually identifiable health information. Federal Register, 65(250), 82462–82829. [PubMed: 11503738]

de Vries S, & Vohra RV (2003). Combinatorial auctions: A survey. INFORMS Journal on Computing, 15(3), 284–309.

Dinev T, & Hart P (2006). An extended privacy calculus model for e-commerce transactions. Information Systems Research, 17(1), 61–80.

Edelman B, Ostrovsky M, & Schwarz M (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. The American Economic Review, 97(1), 242–259.

El Emam K (2013). Guide to the de-identification of personal health information. Boca Raton, FL: CRC Press.

European Parliament and Council of the European Union (EU). (2016, 5 4). General data protection regulation. Official Journal of the European Union. Accessed July 1, 2016, available at http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Federal Trade Commission (FTC). (2014). Data brokers: A call for transparency and accountability. Accessed July 1, 2016, available at http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

Fung BCM, Wang K, Chen R, & Yu PS (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys, 42(4) Article 14, 53 p.

Ghose A, & Yang S (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. Management Science, 55(10), 1605–1622.

Ghosh A, & Roth A (2015). Selling privacy at auction. Games and Economic Behavior, 91, 334–346.

Goldfarb A, & Tucker C (2012). Shifts in privacy concerns. American Economic Review, 102(3), 349–353.

Granville K (2018, 3 19). Facebook and Cambridge Analytica: What you need to know as fallout widens. New York Times. Accessed April 15, 2018, available at https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html

Graeff TR, & Harmon S (2002). Collecting and using personal data: Consumers' awareness and concerns. Journal of Consumer Marketing, 19(4), 302–318.

Grossklags J, & Acquisti A (2007). When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. Proceedings of the Workshop on the Economics of Information Security.

Hallam C, & Zanella G (2017). Online self-disclosure: The privacy paradox explained as a temporally discounted balance between concerns and rewards. Computers in Human Behavior, 68, 217–227.

Hann I-H, Hui K-L, Lee SYT, & Png IPL (2002). Online information privacy: Measuring the cost-benefit trade-off. Proceedings of the 23rd International Conference on Information Systems (pp. 1–10). Atlanta, GA: Association for Information Systems.

Hann IH, Hui K-L, Lee SYT, & Png IPL (2007). Overcoming online information privacy concerns: An information processing theory approach. Journal of Management Information Systems, 24(2), 13–42.

Hastie T, Tibshirani R, & Friedman JH (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

Horowitz JK, & McConnell KE (2002). A review of WTA/WTP studies. Journal of Environmental Economics and Management, 44(3), 426–447.

Hoy MG, & Milne G (2010). Gender differences in privacy-related measures for young adult Facebook users. Journal of Interactive Advertising, 10(2), 28–45.

Huberman BA, Adar E, & Fine LR (2005). Valuating privacy. IEEE Security and Privacy, 3(5), 22–25.

Hui K-L, & Png IPL (2006). The economics of privacy. In Hendershott T (Ed.), Handbooks in information systems (pp. 1–27). Amsterdam: Elsevier.

Janda S, & Fair LL (2004). Exploring consumer concerns related to the internet. Journal of Internet Commerce, 3(1), 1–21.

Jensen C, Potts C, & Jensen C (2005). Privacy practices of internet users: Self-reports versus observed behavior. International Journal of Human-Computer Studies, 63(1–2), 203–227.

John LK, Acquisti A, & Loewenstein G (2011). Strangers on a plane: Context-dependent willingness to divulge sensitive information. Journal of Consumer Research, 37(5), 858–873.

Joinson AN, Paine C, Buchanan T, & Reips U - D. (2008). Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. Computers in Human Behavior, 24(5), 2158–2171.

Joinson AN, Reips U-D, Buchanan T, & Schofield CBP (2010). Privacy, trust, and self-disclosure online. Human-Computer Interaction, 25(1), 1–24.

Korkeila K, Suominen S, Ahvenainen J, Ojanlatva A, Rautava P, Helenius H, & Koskenvuo M (2001). Non-response and related factors in a nationwide health survey. European Journal of Epidemiology, 17(11), 991–999. [PubMed: 12380710]

Krasnova H, Hildebrand T, & Guenther O (2009). Investigating the value of privacy on online social networks: Conjoint analysis. Proceedings of the 30th International Conference on Information Systems. Atlanta, GA: Association for Information Systems. Paper 173.

Krishna V (2010). Auction theory. Amsterdam: Elsevier.

Kumaraguru P, & Cranor LF (2005). Privacy indexes: A survey of Westin's studies. ISRI Technical Report, CMU-ISR-5–138, Carnegie Mellon University, Pittsburg, PA.

La Monica PR (2018, 3 23). Facebook has lost $80 billion in market value since its data scandal. CNN Money. Accessed April 15, 2018, available at http://money.cnn.com/2018/03/27/news/companies/facebook-stock-zuckerberg/index.html

Lahaie S, Pennock DM, Saberi A, & Vohra RV (2007). Sponsored search auctions. In Nisan N, Roughgarden T, Tardos E, & Vazirani VV (Eds.), Algorithmic game theory (Chapter 28, pp. 699–716). Cambridge, UK: Cambridge University Press.

Laric MV, Pitta DA, & Katsanis LP (2009). Consumer concerns for healthcare information privacy: A comparison of U.S. and Canadian perspectives. Research in Healthcare Financial Management, 12(1), 93–111.

Laudon KC (1996). Markets and privacy. Communications of the ACM, 39(9), 92–104.

Li Y (2011). Empirical studies on online information privacy concerns: Literature review and an integrative framework. Communications of the Association for Information Systems, 28(1), 453–496.

Li X-B, & Sarkar S (2011). Protecting privacy against record linkage disclosure: A bounded swapping approach for numeric data. Information Systems Research, 22(4), 774–789.

List JA, & Gallet CA (2001). What experimental protocol influence disparities between actual and hypothetical stated values? Environmental and Resource Economics, 20(3), 241–254.

Liu X, & Li X-B (2016) Acquiring high quality customer data with low cost. In Sugumaran V, Yoon V, & Shaw M (Eds.), E-life: Web-enabled convergence of commerce, work, and social life (Lecture notes in business information processing (Vol. 258, pp. 54–65). Switzerland: Springer.

Lusk JL, & Shogren JF (2007). Experimental auctions: Methods and applications in economic and marketing research. Cambridge, UK: Cambridge University Press.

Madrigal AC (2012, 3 19). How much is your data worth? Mmm, somewhere between half a cent and $1,200. The Atlantic. Accessed July 1, 2016, available at http://www.theatlantic.com/technology/archive/2012/03/how-much-is-your-data-worth-mmm-somewhere-between-half-a-cent-and-1-200/254730

Malhotra NK, Kim SS, & Agarwal J (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information Systems Research, 15(4), 336–355.

McAfee RP, & Vincent D (1993). The declining price anomaly. Journal of Economic Theory, 60, 191–212.

McCambridge J, Witton J, & Elbourne DR (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. Journal of Clinical Epidemiology, 67(3), 267–277. [PubMed: 24275499]

Morey T, Forbath T, & Schoop A (2015). Customer data: Designing for transparency and trust. Harvard Business Review, 93(5), 96–105.

O'Neil D (2001). Analysis of internet users' level of online privacy concerns. Social Science Computer Review, 19(1), 17–31.

Palet LS (2014, 6 30). Ozy.com: Selling your own data. USA Today. Accessed July 1, 2016, available at http://www.usatoday.com/story/money/business/2014/06/30/ozy-selling-data/11760339

Pavlou PA (2011). State of the information privacy literature: Where are we now and where should we go? MIS Quarterly, 35(4), 977–988.

Phelps J, Nowak G, & Ferrell E (2000). Privacy concerns and consumer willingness to provide personal information. Journal of Public Policy and Marketing, 19(1), 27–41.

Ponemon Institute. (2017). 2017 Cost of data breach study, accessed December 20, 2017, available at https://www.ibm.com/security/data-breach/

Preibusch S (2013). Guide to measuring privacy concern: Review of survey and observational instruments. International Journal of Human-Computer Studies, 71(12), 1133–1143.

Premazzi K, Castaldo S, Grosso M, Raman P, Brudvig S, & Hofacker CF (2010). Customer information sharing with e-vendors: The roles of incentives and trust. International Journal of Electronic Commerce, 14(3), 63–91.

Rothkopf MH (2007). Thirteen reasons why the Vickrey-Clarke-Groves process is not practical. Operations Research, 55(2), 191–197.

Rothkopf MH, Pekeč A, & Harstad RM (1998). Computationally manageable combinatorial auctions. Management Science, 44(8), 1131–1147.

Schreiner M, & Hess T (2015). Why are consumers willing to pay for privacy? An application of the privacy-freemium model to media companies. Proceedings of the 2015 European Conference on Information Systems. Atlanta, GA: Association for Information Systems.

Schreiner M, Hess T, & Faranak F (2013). On the willingness to pay for privacy as a freemium model: First empirical evidence. Proceedings of the 2013 European Conference on Information Systems. Atlanta, GA: Association for Information Systems.

Shoemaker PJ, Eichholz M, & Skewes EA (2002). Item nonresponse: Distinguishing between don't know and refuse. International Journal of Public Opinion Research, 14(2), 193–201.

Simonite T (2014, 2 12). Sell your personal data for $8 a month. MIT Technology Review. Accessed July 1, 2016, available at http://www.technologyreview.com/news/524621/sell-your-personal-data-for-8-a-month/

Sledge M (2013, 3 4). Alex Kozinksi, federal judge, would pay a maximum of $2,400 a year for privacy. Huffington Post. Accessed July 1, 2016, available at http://www.huffingtonpost.com/2013/03/04/alex-kozinski-privacy_n_2807608.html

Smith HJ, Dinev T, & Xu H (2011). Information privacy research: An interdisciplinary review. MIS Quarterly, 35(4), 989–1015.

Smith HJ, Milberg JS, & Burke JS (1996). Information privacy: Measuring individuals' concerns about organizational practices. MIS Quarterly, 20(2), 167–196.

Spiekermann S, & Korunovska J (2017). Towards a value theory for personal data. Journal of Information Technology, 32(1), 62–84.

Staiano J, Oliver N, Lepri B, de Oliveira R, Caraviello M, & Sebe N (2014). Money walks: A human-centric study on the economics of personal mobile data. Proceedings of the 2014 ACM Conference on Ubiquitous Computing (pp. 583–594). New York: ACM Press.

Steel E, Locke C, Cadman E, & Freese B (2013, 6 12). How much is your personal data worth? Financial Times. Retrieved from http://www.ft.com/intl/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html

Sweeney L (2002). *k*-Anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 557–570.

Sydell L (2018, 3 26). FTC confirms it's investigating Facebook for possible privacy violations. National Public Radio (NPR). Accessed July 1, 2016, available at https://www.npr.org/sections/thetwo-way/2018/03/26/597135373/ftc-confirms-its-investigating-facebook-for-possible-privacy-violations

Syed R, Dhillon G, & Merrick J (2019). The identity management value model: A design science approach to assess value gaps on social media. Decision Sciences, 50(3), 498–536.

Varian HR (2007). Position auctions. International Journal of Industrial Organization, 25(6), 1163–1178.

Varian H, Wallenberg F, & Woroch G (2005). The demographics of the do-not-call list. IEEE Security and Privacy, 3(1), 34–39.

Warren E, & Warner M (2018, 1 10). Data Breach Prevention and Compensation Act of 2018. Accessed April 15, 2018, available at https://www.congress.gov/bill/115th-congress/senate-bill/2289/text

Wold S, Sjöström M, & Eriksson L (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130.

Woodruff A, Pihur V, Consolvo S, Schmidt L, Brandimarte L, & Acquisti A (2014). Would a privacy fundamentalist sell their DNA for $1000… if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. Proceedings of the 10th Symposium on Usable Privacy and Security, Berkeley, CA: USENIX Association.

Zukowski T, & Brown I(2007). Examining the influence of demographic factors on internet users' information privacy concerns. Proceedings of the 2007 Annual Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries (pp. 197–204). New York: ACM Press.

Table 1:

Types and attributes of personal data.

Type	Identifiers (I)	Quasi-Identifiers (Q)	Demographics (D)	Private (P)
Attributes	Full name Phone number Email address Home address	Date of birth 5-digit zip code	Age Gender Zip code's first three digits Race/Ethnicity Marital status Education (in years) Occupation	Income Credit scores

Table 2:

Summary statistics of participants.

Categorical Attributes	Category	Percentage	Numeric Attributes	Mean	Min.	Max.
Gender	Female	51.8%	Age	32.1	19.0	64.0
	Male	48.2%				
Race/Ethnicity	African American	5.9%	Education (in years)	16.3	12.0	30.0
	Asian	17.1%				
	Hispanic	6.9%	Income ($)	61,922	0	300,000
	Other	2.9%				
	White	67.2%				
Marital Status	Never Married	62.2%				
	Married	28.4%				
	Divorced/Separated	7.4%				
	Other	2.0%				

Table 3:

Average bid prices (in dollars).

		Gender		Race		Marital Status		Student Status	
Data Type	**Overall**	**Female**	**Male**	**White**	**Others**	**Never Married**	**Others**	**Student**	**Others**
Identifier	89.76	90.04	89.45	86.39	81.28	68.46	111.62	28.34	125.18
Quasi-id	88.94	109.49	67.41	90.36	78.60	70.31	112.73	21.67	127.24
Demographic	78.91	90.65	66.38	70.46	77.59	62.17	90.13	25.53	109.46
Private	117.78	130.20	104.53	117.80	91.04	88.99	141.66	27.73	169.33

Table 4:

Percentages of incomplete data and price comparisons.

Data Types	Percentage of Incomplete	Bidding Price	Nonbidding Price
Identifier	12.84%	$79.16	$163.96
Quasi-identifier	8.72%	$84.95	$135.41
Demographic	6.42%	$72.78	$167.79
Private	16.06%	$100.97	$208.24

Table 5:

Correlations of all variables for valuation.

	Price-I	Price-Q	Price-D	Price-P	Age	Gender	Race	Marital	Education	Income
Price-I	1.000									
Price-Q	0.633	1.000								
Price-D	0.704	0.754	1.000							
Price-P	0.665	0.699	0.771	1.000						
Age	0.221	0.291	0.243	0.334	1.000					
Gender	0.002	0.135	0.093	0.073	0.053	1.000				
Race	−0.019	−0.035	0.027	−0.081	−0.298	−0.016	1.000			
Marital	0.166	0.133	0.109	0.165	0.634	0.055	−0.200	1.000		
Education	0.202	0.170	0.116	0.180	0.403	0.041	−0.002	0.239	1.000	
Income	−0.070	0.072	−0.062	−0.068	0.455	−0.242	−0.192	0.235	0.453	1.000

Table 6:

Results of OLS regression model (3).

	SqrtPrice-I	SqrtPrice-Q	SqrtPrice-D	SqrtPrice-P	Price-I	Price-Q	Price-D	Price-P
Constant	2.495	3.112	1.486	3.240	−20.05	34.02	−28.39	−4.02
Age	0.063	0.128***	0.094**	0.158***	1.51	3.49***	2.19**	4.48***
Gender	0.681	1.629**	1.461**	2.099***	11.95	36.53*	30.10**	46.88**
Race	0.331	0.853	1.575**	−0.244	9.64	17.02	31.92*	−3.96
Marital	0.536	−0.091	0.424	−0.334	13.60	−13.50	5.79	−16.64
Education	0.338**	0.189	0.053	0.211	6.78	2.20	−0.13	3.57
Complete	−3.364**	−4.478*	N/A	−4.753***	−73.36**	−123.02*	N/A	−120.44***
R^2	0.108	0.146	0.106	0.186	0.090	0.119	0.088	0.182

Significance level

*** $\alpha = .01$

** $\alpha = .05$

* $\alpha = .10$.

Table 7:

Results of OLS regression model (4).

	SqrtPrice-I	SqrtPrice-Q	SqrtPrice-D	SqrtPrice-P	Price-I	Price-Q	Price-D	Price-P
Constant	−3.092	−4.625	−2.884	−5.721	−140.90	−141.79	−106.06	−190.27*
Age	0.048	0.118***	0.093**	0.134***	1.20	3.29***	2.17**	4.07***
Gender	0.460	1.409*	1.259**	1.702**	7.97	33.46*	26.88*	38.18*
Race	0.694	1.548*	1.992***	0.302	18.55	37.42*	41.39**	10.91
Marital	0.194	−0.424	0.026	−0.875	7.59	−24.07	−0.95	−29.32
Education	0.339**	0.217	0.070	0.219	6.94*	3.28	0.38	4.08
Complete	−3.592***	−3.294	N/A	−4.721***	−78.83**	−88.67	N/A	−118.48***
Fundamental	1.136	2.553	1.126	0.657	22.42	63.26	14.86	7.25
Pragmatist	−0.382	−0.592	−0.734	−1.582	−15.80	−33.22	−29.15	−57.06
ShareInfo	0.358	0.195	0.071	0.642*	6.30	1.29	0.54	11.04
Facebook	0.667*	0.894**	0.727**	1.090***	14.08	19.13*	13.22	25.83**
Twitter	0.840**	0.747*	0.497	1.419***	20.62**	19.60*	11.66	30.75***
R^2	0.158	0.231	0.157	0.287	0.135	0.219	0.135	0.269

Significance level

*** $\alpha = .01$

** $\alpha = .05$

* $\alpha = .10$.

Table 8:

Results of performance evaluation.

	ADF-Gender	ADF-Race	ADF-Marital	ADM-Age	ADM-Education	ADM-Income	RMSE
GSP	3.85%	4.60%	1.86%	0.33%	0.01%	3.83%	15.86
BDM	8.68%	9.29%	2.41%	4.25%	0.52%	9.38%	16.75