

A Supply-Side Pricing Model for Personal Data: Privacy Compensation, Sensitivity, and Information Entropy

Juguang Chen

Guangzhou Foreign Language School, Guangzhou, China

Email: 3338349474@qq.com

How to cite this paper: Chen, J. G. (2026). A Supply-Side Pricing Model for Personal Data: Privacy Compensation, Sensitivity, and Information Entropy. *Theoretical Economics Letters*, 16, 978-998.

<https://doi.org/10.4236/tel.2026.165054>

Received: June 28, 2026

Accepted: October 12, 2026

Published: October 15, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Personal-data markets face a pricing problem because data buyers value information content while data providers bear heterogeneous privacy costs. This paper develops a supply-side reference-pricing framework that combines entropy-based information content, information-category sensitivity weights, and stylized privacy-compensation regimes. A Wenjuanxing survey conducted in Guangdong Province, primarily Guangzhou, from March to May 2024 retained 103 of 130 submissions after documented quality screening. The six-item privacy-compensation preference measure showed acceptable internal consistency (Cronbach's $\alpha = 0.81$; $M = 3.81$, $SD = 0.52$). Exploratory comparisons indicate differences by age, $F(5,97) = 3.884$, $p = 0.003$, and prior privacy-breach experience, $F(2,100) = 5.159$, $p = 0.007$, whereas gender and education are not significant at the 5% level. All respondents rated the full 16-profile orthogonal conjoint set, yielding normalized sensitivity coefficients highest for biometric information (0.190). The entropy component is calculated from a verified 12-profile, 39-attribute matrix (468 binary attribute opportunities) using observed category frequencies. After correcting the compensation schedules so that $L = 7$ is used consistently and every benchmark factor follows Equation (4), a 1500-yuan baseline produces reference prices from approximately 1331 to 3155 yuan. The output is a transparent supply-side negotiation benchmark rather than a market-clearing price.

Keywords

Personal Data Markets, Privacy Compensation, Willingness to Accept, Information Entropy, Data Pricing

1. Introduction

Personal data have become an economically valuable input in digital markets, but

their exchange differs from conventional commodity trade. A single data record can be copied at low marginal cost, combined with other records, and used repeatedly, while disclosure may impose privacy costs on the individual who supplies the information. The resulting valuation problem is therefore two-sided: buyers care about the informational contribution of data, whereas providers care about the expected loss of privacy, control, and future exposure. A workable pricing rule must distinguish these two components rather than treating the value of personal data as a single intrinsic number.

The economics of privacy provides a natural foundation for this problem. Privacy choices involve trade-offs under imperfect information, and the value individuals attach to protecting or disclosing personal information is heterogeneous across people and contexts (Acquisti et al., 2016). Recent work on data markets further shows that personal data may generate informational externalities: one person's disclosure can reveal information about others, which can alter incentives and depress data prices (Acemoglu et al., 2022; Bergemann et al., 2022). Data markets may therefore exist even when consumers are privacy conscious, but the terms of exchange depend on incentives, data linkages, and the way privacy costs are represented (Argenziano & Bonatti, 2023).

Existing data-pricing research has made important progress on query-based prices, arbitrage-free mechanisms, data quality, and information-content measures (Balazinska et al., 2011; Koutris et al., 2012; Lin & Kifer, 2014; Shen et al., 2016; Yu & Zhang, 2017). Yet these approaches often price data from the seller or platform perspective and do not directly incorporate heterogeneous reservation compensation on the supply side. At the same time, empirical privacy-valuation research shows that willingness to accept (WTA) compensation varies by information type and disclosure context; for example, consumers may value personally identifiable and medical information differently from less sensitive information (Liu et al., 2023). This creates a practical gap between technical measures of data content and economic measures of privacy loss. A recent privacy-compensation pricing study likewise emphasizes linking provider privacy costs to data-value components (Dai et al., 2026).

This paper addresses that gap by developing a supply-side reference-pricing model for personal data. The model integrates three elements. First, Shannon information entropy captures the informational content of a record. Second, conjoint-derived sensitivity coefficients distinguish the relative privacy weight of seven information categories. Third, stylized compensation functions represent different degrees of tolerance for privacy disclosure. These elements are combined into a transparent record-level pricing factor and then scaled by a baseline monetary value.

The empirical component is deliberately modest. A survey of 103 respondents is used to describe heterogeneity in privacy-compensation preferences and to calibrate information-category sensitivity. The survey is not treated as a nationally representative estimate of population privacy preferences, and the three compen-

sation functions are not claimed to be latent classes recovered from the sample. Instead, the functions are stylized economic regimes that translate differing levels of privacy tolerance into alternative compensation schedules. This distinction is important because it separates what the data directly support from what is imposed as a modeling device.

The paper makes three contributions. First, it reframes personal-data pricing as a supply-side reservation-compensation problem rather than as a purely technical valuation exercise. Second, it connects an economic measure of privacy compensation with an information-theoretic measure of data content while preserving the distinction between the two. Third, it provides a numerical illustration showing how the same baseline price can generate different reference prices across data profiles when sensitivity, information content, and compensation regime vary. The proposed output should be interpreted as a pricing benchmark for negotiation or platform design, not as an equilibrium market price.

2. Literature and Economic Framework

2.1. Data Pricing and the Economics of Privacy

Data pricing has been studied from several complementary directions. Query-based mechanisms price access to database views while avoiding arbitrage (Koutris et al., 2012; Lin & Kifer, 2014). Other models incorporate data quality, completeness, timeliness, or tuple-level characteristics into the valuation rule (Shen et al., 2016; Yu & Zhang, 2017). Asset-valuation approaches extend conventional valuation concepts to data, while information-theoretic approaches use entropy or related measures to describe scarcity, uncertainty, or information content (Yu et al., 2018). These contributions clarify the technical side of data value but do not fully determine how much a specific individual must be compensated to supply privacy-sensitive data.

From an economic perspective, the provider's reservation value is central. WTA is the minimum compensation that makes disclosure acceptable to the provider. In privacy settings, WTA need not equal a socially efficient price because individuals may face limited information about future data use, data may create externalities, and preferences may be context dependent (Acquisti et al., 2016; Acemoglu et al., 2022). Nevertheless, WTA remains a useful supply-side concept because it makes explicit that privacy loss is a cost borne by the provider rather than a free input to the data buyer.

The present model therefore does not attempt to solve the full equilibrium of a data market. It constructs a reference price from the provider side. Buyer willingness to pay, platform bargaining power, downstream externalities, and competition would still be required to obtain a market-clearing transaction price. This narrower interpretation improves the economic meaning of the model: the output is a structured reservation-compensation benchmark conditioned on the informational contribution of the supplied data.

2.2. Information Entropy and the Model's Information-Content Term

Let x_j denote a category or state and $P(x_j)$ its empirical probability. Shannon self-information for a realized state is $-\log_2 P(x_j)$, while Shannon entropy is the probability-weighted expectation over possible states. The classic entropy form is:

$$H = -\sum_j P(x_j) \log_2 P(x_j) \quad (1)$$

Equation (1) provides the information-theoretic foundation, but the record-level quantity H_i used later is deliberately narrower. Because the pricing exercise assigns value only to attributes that are present and disclosed in a profile, H_i aggregates the present-state Shannon contribution $-P(x_j) \log_2 P(x_j)$, weighted by category sensitivity and attribute count. It is therefore a sensitivity-weighted information-content term derived from Shannon entropy, not the full Shannon entropy of the complete binary vector. For a binary variable, the full entropy would additionally include the absent-state term $-(1-P) \log_2 (1-P)$. This distinction is maintained throughout the revised model.

2.3. Privacy Sensitivity and WTA

Personal information is heterogeneous. Identification data, financial records, medical information, location traces, preferences, communication-network information, and biometrics differ in both perceived sensitivity and economic consequences. The Information Security Technology—Personal Information Security Specification (GB/T 35273-2020) (State Administration for Market Regulation & National Standardization Administration, 2020) provides a practical classification framework, and prior empirical work similarly distinguishes information categories when studying disclosure and privacy concern (Bansal et al., 2010; Chua et al., 2021; Liu et al., 2023).

Table 1. Survey components and their role in the pricing framework.

Component	Content	Role in analysis
Demographics	Gender, age, education, occupation, privacy-breach experience	Describes sample and exploratory heterogeneity
Privacy-compensation preference	Three compensation-attitude items and three disclosure-willingness items on 5-point scales	Exploratory group comparisons; not used as a population risk-preference scale
Conjoint profiles	Compensation amount plus seven personal-information categories	Derives relative utilities and normalized sensitivity coefficients

Note: Privacy-compensation preference is distinguished from domain-general risk attitude to avoid overinterpreting the six survey items.

The survey used in this study combines attitudinal items concerning compensation and disclosure with a conjoint exercise. The attitudinal items are treated as indicators of privacy-compensation preference, not as a validated measure of general risk preference in the behavioral-economics sense. This terminology avoids equating willingness to disclose personal data with domain-general risk aversion. The conjoint component is used to derive normalized sensitivity weights for the seven information categories. These survey components and their analytical roles are summarized in **Table 1**.

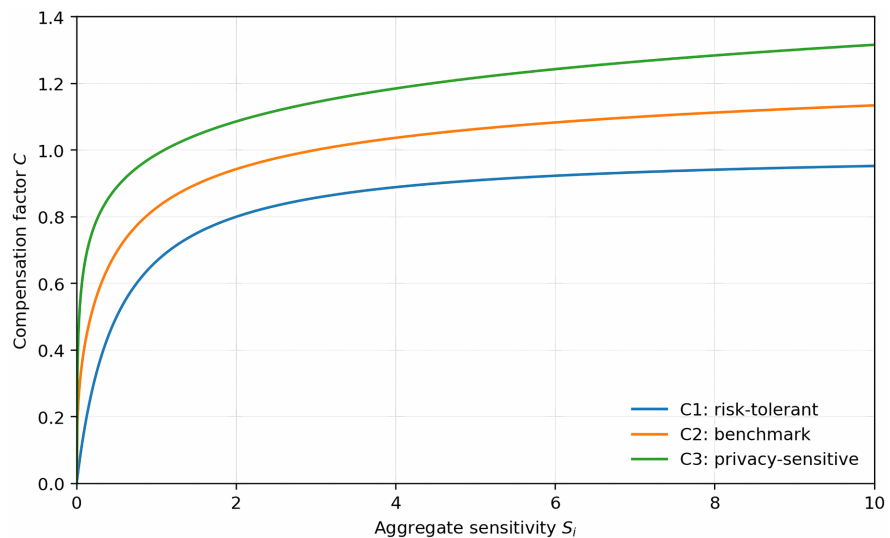
2.4. Stylized Privacy-Compensation Regimes

To represent heterogeneous provider tolerance for privacy disclosure, the model uses three stylized compensation regimes. They are not estimated latent classes. Rather, they provide comparative pricing schedules for a relatively risk-tolerant provider, a benchmark provider, and a privacy-sensitive provider. Let S_i denote the aggregate sensitivity of all information supplied by provider i , and let L denote the number of personal-information categories. In all calculations below, $L = 7$. The three compensation factors are:

$$C_1 = 1 - \frac{1}{1 + 2S_i} \quad (2)$$

$$C_3 = \frac{\ln(10^3 \times S_i + 1)}{L} \quad (3)$$

$$C_2 = \frac{1 - \frac{1}{1 + 2S_i} + \frac{\ln(10^3 \times S_i + 1)}{L}}{2} \quad (4)$$



Note. C_1 represents the relatively risk-tolerant schedule, C_2 the benchmark schedule, and C_3 the privacy-sensitive schedule. The plotted functions use $L = 7$; at $S_i = 10$ they yield $C_1 = 0.952$, $C_2 = 1.134$, and $C_3 = 1.316$.

Figure 1. Stylized privacy-compensation schedules.

The benchmark compensation factor is the average of the relatively risk-tolerant and privacy-sensitive schedules. Aggregate sensitivity is calculated from the quantity of contributed information and the category-specific sensitivity coefficients:

$$S_i = \sum_{j=1}^m T_{ij} S_j \quad (5)$$

These functions are dimensionless and monotone in aggregate sensitivity. Specifically, $dC_1/dS_i = 2/(1+2S_i)^2 > 0$ and $dC_3/dS_i = 1000/[7(1000S_i+1)] > 0$; because $C_2 = (C_1 + C_3)/2$, $dC_2/dS_i > 0$ as well. The benchmark schedule is therefore exactly the midpoint between the two limiting schedules for every S_i . For example, at $S_i = 10$, $C_1 = 0.952$, $C_3 = 1.316$, and $C_2 = 1.134$. The numerical illustration below calculates every benchmark factor from Equation (4), rather than assigning benchmark values independently (Figure 1).

3. Research Design

3.1. Survey Procedure, Documentation, and Sample

The study used an online questionnaire administered through Wenjuanxing between March and May 2024. Participants were recruited by convenience sampling through WeChat and personal networks and through school and workplace channels, with limited onward sharing of the questionnaire link. Recruitment was geographically concentrated in Guangdong Province, primarily Guangzhou. Eligibility required respondents to be able to read Chinese, use Internet or mobile services, and complete the online questionnaire; the study did not impose an age-18 eligibility threshold. A total of 130 submissions were recorded. Data-quality screening excluded 27 cases: 12 were incomplete or missing key questionnaire/conjoint sections, 7 had completion times below one-third of the sample median, 5 showed clear straight-line responding or logical inconsistencies, and 3 were duplicate submissions identified from the platform record and response pattern. The final analytical sample therefore contained 103 respondents, a retained-response rate of 79.2%. Because recruitment was non-probability based, the sample is treated as a calibration sample rather than a population-representative survey.

Privacy-compensation preference was measured with six 5-point items (1 = strongly disagree, 5 = strongly agree). The three compensation-attitude items were: C1, "I should receive compensation when an organization collects or uses my personal information"; C2, "The more sensitive the personal information is, the higher the compensation should be"; and C3, "I would require higher compensation when more categories of my personal information are disclosed." The three disclosure-willingness items were: D1, "I am willing to provide basic identity information when the compensation is adequate"; D2, "I am willing to provide location or behavioral information when the compensation is adequate"; and D3, "I am willing to provide highly sensitive information (e.g., financial, health, or biometric data) when the compensation is adequate." C1-C3 were scored in the original direction and D1-D3 were reverse-coded so that higher values consist-

ently indicated a stronger compensation requirement. The composite score was the mean of the six items after reverse coding. Internal consistency was Cronbach's $\alpha = 0.81$; the composite had $M = 3.81$ and $SD = 0.52$ in the 103 retained cases. The measure is interpreted specifically as privacy-compensation preference rather than as a domain-general risk-aversion scale.

Within the retained sample, men account for 52.43% and women 47.57%. The age distribution is concentrated in the 36 - 55 range: respondents aged 46 - 55 constitute 46.60% and those aged 36 - 45 constitute 30.10%. Bachelor's degree holders are the largest education group at 33.98%, followed by associate-degree/junior-college respondents at 25.24%. Corporate employees account for 35.92% of the sample. Ninety-three respondents (90.29%) report that they regard privacy as very important. All sample counts and analytical inputs reported below were cross-checked against the archived questionnaire export, screening log, conjoint record, and calculation files before the revised model was recomputed. Detailed sample characteristics are reported in **Table 2**.

Table 2. Sample characteristics (N = 103).

Characteristic	Category	n	%
Gender	Male	54	52.43
Gender	Female	49	47.57
Age	Below 18	7	6.80
Age	18 - 25	3	2.91
Age	26 - 35	7	6.80
Age	36 - 45	31	30.10
Age	46 - 55	48	46.60
Age	56 or above	7	6.80
Education	High school or below	19	18.45
Education	Associate/junior college	26	25.24
Education	Bachelor's degree	35	33.98
Education	Master's degree	18	17.48
Education	Doctoral degree or higher	5	4.85
Occupation	Corporate employee	37	35.92
Occupation	Freelancer	23	22.33
Occupation	Civil servant/public institution	12	11.65
Occupation	Student	8	7.77
Occupation	Other	23	22.33
Privacy importance	Very important	93	90.29

3.2. Personal-Information Categories

Drawing on prior classifications and GB/T 35273-2020, the study organizes personal information into seven categories. The categories are broad enough to support a parsimonious pricing rule while retaining economically meaningful differences in privacy sensitivity. The seven personal-information categories and their illustrative content are presented in **Table 3**.

Table 3. Personal-information categories.

Category	Illustrative content
Basic identity	ID number, name, age, gender, marital status, occupation, education, residential information
Communication network	Phone number, email, account identifiers, network-use records, IP address, communication records
Financial	Credit information, income, consumption records, bank-account details, assets
Preference	Interests, traits, preferences, habits, religious beliefs, political leanings
Location	Movement trajectories and precise positioning
Medical and health	Medical visits, medical history, health status
Biometric	Genes, fingerprints, voice, palm prints, iris and facial-recognition features

Note: Classification is based on the study's original coding scheme and GB/T 35273-2020 (State Administration for Market Regulation & National Standardization Administration, 2020).

3.3. Conjoint Design and Sensitivity Calibration

The conjoint task treated each hypothetical data-sharing profile as a bundle of attributes. The compensation attribute had three levels-300, 1500, and 3000 yuan-and each of the seven personal-information categories was coded as present or absent. An orthogonal design reduced the full factorial space to 16 profiles. Every respondent evaluated all 16 profiles on a 5-point acceptance/disclosure-willingness rating scale (1 = very unwilling, 5 = very willing). The orthogonal profile set was fixed, while Wenjuanxing randomized the order in which profiles were presented to respondents to reduce systematic order effects.

The monetary amounts are experimental levels rather than externally validated transaction prices. Following standard additive conjoint practice (Green & Srinivasan, 1990; Luce & Tukey, 1964), an additive conjoint model was estimated across the 16 ratings. For information category j , its utility range was calculated as $r_j = \max(u_{ji}) - \min(u_{ji})$ across the present/absent levels; category sensitivity was then normalized as $s_j = r_j / \sum_{k=1}^7 r_k$, so that the seven information-category

coefficients sum to one. The compensation attribute was estimated jointly in the conjoint model but was not included in this seven-category normalization. Model fit was Pearson's $R = 0.847$ and Kendall's $\tau = 0.712$, indicating reasonably strong agreement between predicted and observed profile ordering within the sample. The conjoint attributes and levels are summarized in **Table 4**.

Table 4. Conjoint attributes and levels.

Attribute	Levels
Compensation amount	300/1500/3000 yuan
Basic identity information	Present/absent
Communication-network information	Present/absent
Financial information	Present/absent
Preference information	Present/absent
Location information	Present/absent
Medical and health information	Present/absent
Biometric information	Present/absent

3.4. Analytical Strategy

The empirical analysis has two linked purposes. First, the six-item privacy-compensation preference score is used in exploratory subgroup comparisons. Gender is tested with an independent-samples t test; age, education, and prior privacy-breach experience are examined with one-way ANOVA. Effect sizes are reported as Cohen's d for gender and eta-squared for the ANOVAs. These subgroup tests describe heterogeneity but do not define the three stylized compensation regimes and are not inputs to Equations (2)-(8). Second, conjoint utility ranges are normalized across the seven information categories to obtain s_j , after which the pricing model combines those category weights with profile-specific information quantities, empirically reconstructed entropy probabilities, and one of the three compensation schedules.

Before group comparisons, score distributions and residual plots were inspected and homogeneity of variance was checked with Levene tests; no material variance violation was detected (all $p > 0.05$). The six pre-specified age categories were retained rather than pooled post hoc, even though the 18 - 25 group contained only three respondents and three additional age cells contained seven respondents. Consequently, the age result is treated as exploratory and no pairwise claim is made from the sparse cells. The doctoral education group ($n = 5$) is handled similarly. The verified statistics are $t(101) = -0.676$, $p = 0.501$ for gender; $F(5,97) = 3.884$, $p = 0.003$ for age; $F(4,98) = 2.396$, $p = 0.055$ for education; and $F(2,100) = 5.159$, $p = 0.007$ for prior privacy-breach experience.

4. Empirical Results

4.1. Exploratory Heterogeneity in Privacy-Compensation Preference

The verified subgroup comparisons are summarized in **Table 5**. Men ($n = 54$, $M = 3.77$, $SD = 0.525$) and women ($n = 49$, $M = 3.84$, $SD = 0.525$) differ only trivially, $t(101) = -0.676$, $p = 0.501$, $d = -0.133$. Across the six age groups, means increase from 3.35 (<18), 3.30 (18 - 25), 3.55 (26 - 35), and 3.72 (36 - 45) to 3.95 (46 - 55) and 4.10 (56+), with a common within-group SD of approximately 0.482; the omnibus difference is $F(5,97) = 3.884$, $p = 0.003$, eta-squared = 0.167. Education means are 3.60, 3.72, 3.86, 4.00, and 4.15 from the lowest to highest category (SD approximately 0.496), but the omnibus result does not reach the 5% threshold, $F(4,98) = 2.396$, $p = 0.055$, eta-squared = 0.089. Prior privacy-breach experience shows a significant difference: no breach ($n = 58$, $M = 3.68$), unsure/suspected ($n = 18$, $M = 3.84$), and yes ($n = 27$, $M = 4.05$), with SD approximately 0.497, $F(2,100) = 5.159$, $p = 0.007$, eta-squared = 0.094. These patterns support heterogeneity in privacy-compensation preference but do not justify mechanically assigning demographic groups to the model's stylized compensation regimes.

Table 5. Exploratory group comparisons in privacy-compensation preference.

Variable	Group sizes	Test (df), p	Effect size	Assumption/handling
Gender	Male 54; Female 49	$t(101) = -0.676$; $p = 0.501$	$d = -0.133$	Levene $p > 0.05$; assumptions acceptable
Age	7/3/7/31/48/7	$F(5,97) = 3.884$; $p = 0.003$	$\eta^2 = 0.167$	All six groups retained; sparse cells; omnibus exploratory; no pairwise claims
Education	19/26/35/18/5	$F(4,98) = 2.396$; $p = 0.055$	$\eta^2 = 0.089$	Levene $p > 0.05$; doctoral $n = 5$; no pairwise claims
Prior privacy breach	No 58; Unsure 18; Yes 27	$F(2,100) = 5.159$; $p = 0.007$	$\eta^2 = 0.094$	Levene $p > 0.05$; assumptions acceptable

Note. All group counts, test statistics, and effect sizes were verified against the archived analysis. The six age categories and five education categories were retained as pre-specified groups; because some cells are small, no pairwise inference is based on the sparse groups.

4.2. Information-Sensitivity Coefficients

The normalized information-category sensitivity coefficients derived from the conjoint utility ranges sum to one. Biometric information receives the largest weight (0.190), followed by financial information (0.141), medical and health information (0.140), basic identity information (0.140), preference information (0.136), location information (0.127), and communication-network information (0.126). These coefficients are sample-specific relative sensitivity weights: they indicate how strongly information categories differentiate disclosure utility within this conjoint calibration, not universal monetary values or legal damage weights. The resulting coefficients are reported in **Table 6**.

Table 6. Normalized personal-information sensitivity coefficients.

Information category	Sensitivity coefficient (s_j)
Basic identity information	0.140
Communication-network information	0.126
Financial information	0.141
Preference information	0.136
Location information	0.127
Medical and health information	0.140
Biometric information	0.190

Note. Coefficients are normalized to sum to 1.000.

5. Integrated Pricing Model and Numerical Illustration

5.1. Record-Level Pricing Factor

Let T_{ij} denote the number of present attributes from category j in profile i , s_j the normalized sensitivity coefficient for category j , and $S_{ij} = T_{ij}s_j$ the category-specific sensitivity contribution. Let $P(x_{ij})$ denote the empirical presence probability for category j and $C(x_i)$ the compensation factor associated with the selected privacy-compensation regime. The model-specific sensitivity-weighted information-content term is $H_i = -\sum_j [T_{ij}s_j P(x_{ij}) \log_2 P(x_{ij})]$, and the record-level pricing factor is $F(x_i) = C(x_i)H_i$. H_i is derived from the present-state Shannon entropy contribution; it is not the full binary Shannon entropy because absent states are intentionally excluded from the pricing index.

$$F(X_i) = -C(x_i) \left(\sum_{j=1}^m S_{ij} P(x_{ij}) \times \log_2 P(x_{ij}) \right) \quad (6)$$

The record-level monetary reference price scales this factor by a baseline value $V_0(x_i)$:

$$V(x_i) = -V_0(x_i) \times C(x_i) \times \left(\sum_{j=1}^m S_{ij} \times P(x_{ij}) \times \log_2 P(x_{ij}) \right) \quad (7)$$

Aggregating across n providers yields the dataset-level supply-side reference value:

$$V = -\sum_{i=1}^m V_0(x_i) C(x_i) \left(\sum_{j=1}^m S_{ij} P(x_{ij}) \times \log_2 P(x_{ij}) \right) \quad (8)$$

Equations (6)-(8) make the units and economic structure explicit. H_i and $C(x_i)$ are dimensionless, while the baseline $V_0(x_i)$ carries the monetary unit; therefore $V(x_i)$ has the same currency unit as $V_0(x_i)$. In the verified numerical matrix below, $P(x_{ij})$ is computed from the observed presence frequency of attributes in the same information category across the 12 profiles. Absent attributes do not enter T_{ij} because the pricing rule values disclosed/present attributes only. Thus H_i should be interpreted as a sensitivity-weighted information-content

index, not as the entropy of the full 39-dimensional binary vector. The model is multiplicative: holding other terms fixed, a larger compensation factor, information-content term, or baseline raises the reference price proportionally.

5.2. Comparative Statics and Interpretation

First, holding the baseline price and information content fixed, a larger compensation factor increases the model-implied reference price proportionally. The three stylized regimes therefore create an explicit privacy-compensation wedge between otherwise identical data profiles. Second, holding the compensation regime fixed, information categories with larger sensitivity coefficients contribute more to the weighted information-content term. Third, because the baseline price enters multiplicatively, changing V_0 rescales all model-implied prices without changing the ranking of profiles, provided the same baseline is used for all observations.

These properties also clarify what the model does not measure. The Shannon-derived information-content term is not buyer willingness to pay, the sensitivity coefficient is not a legal damages schedule, and $C(x_i)$ is not a statistically estimated coefficient from a structural utility model. The framework is therefore best used as a transparent supply-side pricing rule or negotiation benchmark. A complete market model would additionally require demand-side value, bargaining, data externalities, platform competition, and transaction costs.

5.3. Verified Numerical Illustration and Information-Content Inputs

The numerical illustration uses a study-generated 12-profile binary attribute matrix rather than an external public dataset. The matrix contains 39 attributes grouped into seven categories: 9 basic-identity attributes, 3 communication-network attributes, 5 financial attributes, 14 preference attributes, 2 location attributes, 4 medical/health attributes, and 2 biometric attributes. The observation unit for the entropy frequencies is an attribute state within a profile. With 12 profiles and 39 attributes, the matrix contains 468 binary attribute opportunities. Across these opportunities, 304 attributes are present.

Category-specific presence probabilities are calculated directly from the 12-profile matrix: basic identity $87/108 = 0.8056$; communication network $22/36 = 0.6111$; financial $35/60 = 0.5833$; preference $97/168 = 0.5774$; location $15/24 = 0.6250$; medical/health $28/48 = 0.5833$; and biometric $20/24 = 0.8333$. For each present attribute in category j , $P(x_{ij})$ is set to the corresponding category probability. A common baseline $V_0 = 1500$ yuan is used for all 12 profiles. The profile count vector T_{ij} is combined with the sensitivity coefficients reported in **Table 6** to obtain S_i and H_i ; risk-tolerant profiles use C_1 , benchmark profiles use $C_2 = (C_1 + C_3)/2$, and privacy-sensitive profiles use C_3 . Supplementary **Tables S1-S4** provide the complete 12×39 binary realization, category-frequency inputs, and reproducibility calculations. The profile-level information counts and assigned regimes are summarized in **Table 7**.

Table 7. Data profiles in the verified numerical illustration.

No.	ID	Comm.	Fin.	Pref.	Loc.	Health	Bio.	Total	Regime
1	7	1	4	6	2	0	2	22	Benchmark
2	8	2	0	9	1	0	2	22	Privacy-sensitive
3	9	0	2	1	1	4	2	19	Benchmark
4	7	2	1	9	1	3	1	24	Privacy-sensitive
5	7	3	5	12	0	3	2	32	Privacy-sensitive
6	5	1	5	11	2	2	2	28	Benchmark
7	9	2	4	14	1	2	1	33	Benchmark
8	8	2	4	4	2	2	2	24	Risk-tolerant
9	8	3	2	8	1	4	1	27	Risk-tolerant
10	7	1	3	10	0	2	1	24	Benchmark
11	8	2	3	9	2	2	2	28	Benchmark
12	4	3	2	4	2	4	2	21	Benchmark

Note. The 12 rows form the study-generated 12×39 profile matrix used for the numerical illustration. Regime labels are scenario assignments for applying C_1 , C_2 , or C_3 ; they are not demographic classifications.

Table 8. Recomputed pricing results from the verified profile matrix ($V_0 = 1500$ yuan).

No.	S_i	Regime	C_i	Information-content H_i	Reference price (yuan)
1	3.120	Benchmark	1.006	1.121	1691.112
2	3.103	Privacy-sensitive	1.149	1.088	1874.536
3	2.745	Benchmark	0.989	0.898	1331.387
4	3.334	Privacy-sensitive	1.159	1.266	2200.077
5	4.495	Privacy-sensitive	1.202	1.751	3155.288
6	3.941	Benchmark	1.035	1.553	2410.939
7	4.577	Benchmark	1.053	1.775	2803.900
8	3.394	Risk-tolerant	0.872	1.214	1586.564
9	3.745	Risk-tolerant	0.882	1.421	1880.111
10	3.359	Benchmark	1.015	1.284	1954.850
11	3.933	Benchmark	1.035	1.461	2267.381
12	2.958	Benchmark	0.999	1.127	1687.555

The information-content term and compensation factor are recomputed from the documented inputs, as shown in **Table 8**. The revision also corrects the earlier benchmark inconsistency: for profile 1, $S_i = 3.120$ gives $C_1 = 0.862$ and $C_3 = 1.149$, so Equation (4) yields $C_2 = 1.006$. Using $P(x_{ij})$ from the verified category frequencies and $V_0 = 1500$ yuan, the weighted information-content terms range from 0.898 to 1.775 and the resulting reference prices range from 1331.387 yuan (profile 3) to 3155.288 yuan (profile 5). Calculations use unrounded probabilities and compensation factors; displayed values are rounded for readability.

6. Discussion

6.1. Economic Contribution

The central contribution of the framework is conceptual separation. Personal-data pricing contains at least three distinct objects that are often conflated: information content, privacy sensitivity, and provider reservation compensation. Shannon entropy supplies the information-theoretic foundation, while the model-specific H_i is a sensitivity-weighted present-state information-content term; sensitivity weights capture the relative privacy importance of information categories; and the compensation factor represents the provider-side adjustment required under different privacy-tolerance regimes. Multiplying these components creates a tractable reference-pricing rule without claiming that H_i is the full entropy of a binary record or that any single component is itself the market price.

This separation connects the paper to the broader economics of data markets. Prior research emphasizes that data have externalities and that individual disclosure can affect the information available about others (Acemoglu et al., 2022; Bergemann et al., 2022). The present model does not solve those externalities, but it contributes a micro-level pricing primitive that can be embedded in richer market settings: the amount a platform would need to compensate a provider before demand-side value and external effects are considered. In that sense, the model is closer to a reservation-compensation schedule than to a complete theory of data-market equilibrium.

Behavioral evidence remains secondary to the pricing model but is now fully documented. The six-item measure has acceptable internal consistency, and the corrected age and privacy-breach ANOVAs indicate meaningful heterogeneity in privacy-compensation preference. At the same time, the very small age cells, the non-significant gender result, and the education result at $p = 0.055$ argue against demographic pricing rules. The three compensation regimes therefore remain transparent modeling scenarios rather than respondent classes inferred from age, education, gender, or breach history.

6.2. Implications for Data-Market Design

For data platforms, the model provides a transparent way to structure a posted compensation schedule. A platform could begin with a common baseline value, apply category-specific sensitivity weights, and then offer alternative compensa-

tion schedules reflecting consent conditions or privacy guarantees. The approach is especially useful when a platform lacks a long history of comparable transactions and therefore cannot estimate market prices directly.

The framework also illustrates why a single uniform price for all personal data is difficult to justify. Biometric and financial information receive higher sensitivity weights than communication-network and location information in the present calibration, and privacy-sensitive providers receive a larger compensation adjustment than risk-tolerant providers. Even if a platform ultimately uses a simpler commercial rule, these differences identify the dimensions along which a uniform price can undercompensate some providers and overcompensate others.

Policy applications require greater caution. The sensitivity coefficients in this study are sample-specific and should not be used as statutory compensation standards for data breaches. Legal damages depend on rights, harm, causation, and jurisdiction, whereas the present coefficients reflect relative disclosure sensitivity in a survey exercise. The model is therefore more appropriate for voluntary data exchange and platform pricing than for legal damage assessment.

6.3. Limitations and Future Research

Several limitations are important. First, the survey is a small non-probability convenience sample recruited primarily in Guangzhou and the wider Guangdong context. Although recruitment procedures and exclusions are now documented, the sample should not be treated as representative of the broader Chinese population. Replication with a larger probability-based sample would improve the stability and generalizability of subgroup and sensitivity estimates.

Second, the six-item privacy-compensation preference measure is study-specific. Its internal consistency is acceptable ($\alpha = 0.81$), but it is not a validated domain-general risk-preference instrument. Future work should combine privacy-specific WTA items with established behavioral risk measures and test convergent, discriminant, and predictive validity.

Third, the conjoint task is hypothetical. Every respondent rated all 16 profiles and profile order was randomized, which improves internal consistency of administration, but hypothetical ratings may still differ from incentive-compatible disclosure choices. Future studies should use real or consequential compensation where feasible and test whether sensitivity weights replicate out of sample.

Fourth, the category probabilities used in the Shannon-derived information-content term are estimated from the study-generated 12×39 profile matrix rather than from the empirical distribution of a large external data market. This makes the calculation reproducible but still illustrative. Future research should estimate $P(x_{ij})$ from a named external dataset and compare model-implied prices with actual negotiated prices, buyer willingness to pay, or experimentally elicited willingness to accept in a holdout sample.

Fifth, the framework is intentionally supply-side. It does not model informational externalities, bargaining, platform competition, repeated use of data, or the

marginal value of data to a specific buyer. These mechanisms are central to welfare and equilibrium pricing in data markets (Acemoglu et al., 2022; Bergemann et al., 2022). A natural extension is to embed the current reservation-compensation rule in a two-sided bargaining or mechanism-design model.

7. Conclusion

This paper develops a supply-side reference-pricing model for personal data that integrates privacy compensation, information sensitivity, and information entropy. The verified survey data show acceptable reliability for the six-item privacy-compensation preference measure and exploratory heterogeneity by age and prior privacy-breach experience. The conjoint task provides normalized sensitivity weights across seven categories of personal information, with biometric information receiving the largest weight in this sample. These behavioral results inform the calibration but do not mechanically assign individuals to the three stylized compensation regimes.

The model's main value lies in its transparent and reproducible structure rather than in a claim of universal pricing accuracy. The revised implementation uses $L = 7$ consistently, calculates every benchmark factor as the arithmetic mean specified by Equation (4), documents the 12×39 attribute matrix and the empirical category frequencies used for $P(x_{ij})$, and recomputes every information-content term and reference price from those inputs. The accompanying Supplementary Material provides the complete binary realization and step-by-step reproducibility tables. The numerical illustration is therefore internally reproducible, while external market validation remains a task for future research. Framed in this way, the model provides a supply-side negotiation benchmark while preserving the distinction among privacy cost, information content, and market value.

Supplementary Material

Supplementary **Tables S1-S4** report the complete 12×39 binary realization, category-level probability inputs, profile counts, and step-by-step pricing calculations used to reproduce the numerical illustration.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

- Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2022). Too Much Data: Prices and Inefficiencies in Data Markets. *American Economic Journal: Microeconomics*, 14, 218-256. <https://doi.org/10.1257/mic.20200200>
- Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, 54, 442-492. <https://doi.org/10.1257/jel.54.2.442>
- Argenziano, R., & Bonatti, A. (2023). Data Markets with Privacy-Conscious Consumers. *AEA Papers and Proceedings*, 113, 191-196. <https://doi.org/10.1257/pandp.20231083>

- Balazinska, M., Howe, B., & Suciu, D. (2011). Data Markets in the Cloud: An Opportunity for the Database Community. *Proceedings of the VLDB Endowment*, 4, 1482-1485. <https://doi.org/10.14778/3402755.3402801>
- Bansal, G., Zahedi, F. M., & Gefen, D. (2010). The Impact of Personal Dispositions on Information Sensitivity, Privacy Concern and Trust in Disclosing Health Information Online. *Decision Support Systems*, 49, 138-150. <https://doi.org/10.1016/j.dss.2010.01.010>
- Bergemann, D., Bonatti, A., & Gan, T. (2022). The Economics of Social Data. *The RAND Journal of Economics*, 53, 263-296. <https://doi.org/10.1111/1756-2171.12407>
- Chua, H. N., Ooi, J. S., & Herbrand, A. (2021). The Effects of Different Personal Data Categories on Information Privacy Concern and Disclosure. *Computers & Security*, 110, Article ID: 102453. <https://doi.org/10.1016/j.cose.2021.102453>
- Dai, J., Li, S., Liao, Z. et al. (2026). Research on Data Pricing Model Based on Privacy Compensation. *Business and Management*, No. 2, 10-18. (In Chinese)
- Green, P. E., & Srinivasan, V. (1990). Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing*, 54, 3-19. <https://doi.org/10.1177/002224299005400402>
- Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2012). Query-Based Data Pricing. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (pp. 167-178). ACM. <https://doi.org/10.1145/2213556.2213582>
- Lin, B. R., & Kifer, D. (2014). On Arbitrage-Free Pricing for General Data Queries. *Proceedings of the VLDB Endowment*, 7, 757-768. <https://doi.org/10.14778/2732939.2732948>
- Liu, H., Li, K., Chen, Y., & Luo, X. (2023). Is Personally Identifiable Information Really More Valuable? Evidence from Consumers' Willingness-to-Accept Valuation of Their Privacy Information. *Decision Support Systems*, 173, Article ID: 114010. <https://doi.org/10.1016/j.dss.2023.114010>
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology*, 1, 1-27. [https://doi.org/10.1016/0022-2496\(64\)90015-x](https://doi.org/10.1016/0022-2496(64)90015-x)
- Shen, Y., Guo, B., Shen, Y., Duan, X., & Hong, Y. (2016). A Pricing Model for Big Personal Data. *Tsinghua Science and Technology*, 21, 482-490. <https://doi.org/10.1109/tst.2016.7590317>
- State Administration for Market Regulation, & National Standardization Administration (2020). *Information Security Technology-Personal Information Security Specification (GB/T 35273-2020)*. China Standard Press.
- Yu, H., & Zhang, M. (2017). Data Pricing Strategy Based on Data Quality. *Computers & Industrial Engineering*, 112, 1-10. <https://doi.org/10.1016/j.cie.2017.08.008>
- Yu, Y. H., Fu, Y., & Wu, X. P. (2018). Metric and Classification Model for Privacy Data Based on Shannon Information Entropy and BP Neural Network. *Journal on Communications*, 39, 10-17. (In Chinese)

Supplementary Material

S1. Purpose and Interpretation of the Binary Matrix

The numerical illustration is based on 12 profiles and 39 binary attribute indicators grouped into seven information categories. The model enters these data through category-level present counts T_{ij} and category-specific probabilities $P(x_{ij})$. Accordingly, the identity of a particular indicator within a category is exchangeable for the calculation: any within-category permutation leaves T_{ij} , $P(x_{ij})$, S_i , H_i , and the resulting reference price unchanged. **Tables S1(a)-(c)** provide the complete canonical 12×39 binary realization used for reproducibility, with 1 indicating a present/disclosed attribute and 0 indicating an absent attribute. The matrix exactly reproduces the verified category counts reported in the manuscript.

Table S1. (a) Binary matrix: basic identity and communication network; (b) Binary matrix: financial and preference indicators 1 - 7; (c) Binary matrix: preference indicators 8 - 14, location, medical/health, and biometric.

(a)												
Profile	ID01	ID02	ID03	ID04	ID05	ID06	ID07	ID08	ID09	COM01	COM02	COM03
1	1	1	1	1	1	1	1	0	0	1	0	0
2	1	1	1	1	1	1	1	0	1	1	0	1
3	1	1	1	1	1	1	1	1	1	0	0	0
4	1	1	1	1	1	1	1	0	0	1	1	0
5	1	1	1	1	1	1	0	0	1	1	1	1
6	1	1	1	0	0	0	0	1	1	0	1	0
7	1	1	1	1	1	1	1	1	1	1	1	0
8	1	1	1	1	0	1	1	1	1	1	1	0
9	1	1	1	1	0	1	1	1	1	1	1	1
10	1	1	1	0	0	1	1	1	1	0	1	0
11	1	1	1	0	1	1	1	1	1	1	1	0
12	0	0	0	0	1	1	1	1	0	1	1	1

(b)												
Profile	FIN01	FIN02	FIN03	FIN04	FIN05	PREF01	PREF02	PREF03	PREF04	PREF05	PREF06	PREF07
1	1	1	1	1	0	1	1	1	1	1	1	0
2	0	0	0	0	0	1	1	0	0	0	0	0
3	0	1	1	0	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0	0	0	1	1
5	1	1	1	1	1	0	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1

Continued

7	1	1	0	1	1	1	1	1	1	1	1	1	1
8	1	1	0	1	1	1	1	1	0	0	0	0	0
9	0	0	0	1	1	0	0	0	0	1	1	1	1
10	0	1	1	1	0	1	1	1	1	1	1	1	1
11	1	1	1	0	0	1	1	1	1	1	0	0	0
12	1	0	0	0	1	0	0	0	0	0	0	0	1

(c)

Profile	PREF 08	PREF 09	PREF 10	PREF 11	PREF 12	PREF 13	PREF 14	LOC 01	LOC 02	HLT 01	HLT 02	HLT 03	HLT 04	BIO 01	BIO 02
1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1
2	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1
3	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1
5	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1
6	1	1	1	1	0	0	0	1	1	1	0	0	1	1	1
7	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1
8	0	0	0	0	0	0	1	1	1	0	1	1	0	1	1
9	1	1	1	1	1	0	0	1	0	1	1	1	1	1	0
10	1	1	0	0	0	0	1	0	0	0	1	1	0	1	0
11	0	0	0	1	1	1	1	1	1	1	1	0	0	1	1
12	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1

Note. Column codes identify binary indicators within each information category; only category membership and the present/absent state enter the pricing calculation.

S2. Verified Profile-Level Category Counts

Table S2 reproduces the category count vector T_{ij} used in the main-text numerical illustration and verifies that each row of **Tables S1(a)-(c)** yields the same profile totals.

Table S2. Profile counts and compensation regimes.

Profile	ID	Comm.	Fin.	Pref.	Loc.	Health	Bio.	Total	Regime
1	7	1	4	6	2	0	2	22	Benchmark
2	8	2	0	9	1	0	2	22	Privacy-sensitive
3	9	0	2	1	1	4	2	19	Benchmark
4	7	2	1	9	1	3	1	24	Privacy-sensitive

Continued

5	7	3	5	12	0	3	2	32	Privacy-sensitive
6	5	1	5	11	2	2	2	28	Benchmark
7	9	2	4	14	1	2	1	33	Benchmark
8	8	2	4	4	2	2	2	24	Risk-tolerant
9	8	3	2	8	1	4	1	27	Risk-tolerant
10	7	1	3	10	0	2	1	24	Benchmark
11	8	2	3	9	2	2	2	28	Benchmark
12	4	3	2	4	2	4	2	21	Benchmark

S3. Category Frequencies and Probability Inputs

For category j , the presence probability p_j is the number of present binary states divided by the number of possible states across the 12 profiles. These probabilities supply the present-state Shannon contribution $-p_j \log_2(p_j)$ used in H_i .

Table S3. Frequency inputs used to calculate $P(x_{ij})$.

Category	Attributes/profile	Present count	Opportunities	$P(x_{ij})$	$-p \log_2 p$
Basic identity	9	87	108	0.8056	0.2513
Communication network	3	22	36	0.6111	0.4342
Financial	5	35	60	0.5833	0.4536
Preference	14	97	168	0.5774	0.4575
Location	2	15	24	0.6250	0.4238
Medical and health	4	28	48	0.5833	0.4536
Biometric	2	20	24	0.8333	0.2192
Total	39	304	468	0.6496 overall	—

S4. Clarification of H_i and Reproducibility Formula

For a binary variable with presence probability p , full Shannon entropy is $H_B(p) = -p \log_2(p) - (1-p) \log_2(1-p)$. The pricing model does not use $H_B(p)$ as the record-level quantity because the monetary rule is applied only to attributes that are present and disclosed in a profile. Instead, it uses the present-state entropy contribution $-p \log_2(p)$, multiplied by the number of present attributes T_{ij} and the category sensitivity weight s_j . Thus $H_i = -\sum_j T_{ij} s_j p_j \log_2(p_j)$. H_i should therefore be read as a sensitivity-weighted information-content term derived from Shannon entropy, not as the full binary Shannon entropy of the 39-dimensional profile. This terminology change clarifies the information-theoretic interpretation without changing any calculation.

Table S4. Reproducibility check for profile-level pricing outputs.

Profile	S_i	Regime	C_i	H_i	V_0 (yuan)	Reference price (yuan)
1	3.120	Benchmark	1.006	1.121	1500	1691.112
2	3.103	Privacy-sensitive	1.149	1.088	1500	1874.536
3	2.745	Benchmark	0.989	0.898	1500	1331.387
4	3.334	Privacy-sensitive	1.159	1.266	1500	2200.077
5	4.495	Privacy-sensitive	1.202	1.751	1500	3155.288
6	3.941	Benchmark	1.035	1.553	1500	2410.939
7	4.577	Benchmark	1.053	1.775	1500	2803.900
8	3.394	Risk-tolerant	0.872	1.214	1500	1586.564
9	3.745	Risk-tolerant	0.882	1.421	1500	1880.111
10	3.359	Benchmark	1.015	1.284	1500	1954.850
11	3.933	Benchmark	1.035	1.461	1500	2267.381
12	2.958	Benchmark	0.999	1.127	1500	1687.555

Reproducibility note. Calculations in the manuscript use unrounded category probabilities and compensation factors; displayed values are rounded for readability. Because within-category indicators share the same s_j and p_j , within-category column permutations in **Table S1** do not alter the reported H_i or reference prices.