For office use only	Team Control Number	For office use only
T1 _____	**93036**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**F**	F4 _____

2018
MCM/ICM
Summary Sheet

PIPE: Estimate the Value of Private Information

Summary

Contrary to the pervasive belief that human society has entered the information age, the massive data produced by human individuals are not fully exploited yet. Private data nowadays are under poor, isolated management by individual enterprises, where the value of data cannot be fully extracted to benefit either its provider or owner. To address this problem, a well-established market system is required that not only prices and rewards data sharing, but also regulates and protects private information.

To satisfy the requirement, our paper provides a detailed analysis based on a dataset *PI-DATA*, based on which we propose a sophisticated and generalized model, Private Information Price Estimation (PIPE), which is able to estimate the price of private information (PI) regarding different data domains of PI and social subgroups.

Task 1: We abstractly extract *feature vectors* from individuals and query requests to distinctly characterize their traits in different data categories..

Task 2: We estimate the correlation matrix of data categories and develop an amendment formula to accurately compute data value considering internal and external factors.

Task 3: We establish a *Supply and Demand Model* to estimate the value of PI as a commodity on the level of individuals, groups and nations.

Task 4: We surveyed the existing government act (e.g. Privacy Act, GDPR, APPI, etc.) and price regulations related to the private information around the world. Also, we introduce a dynamic variation to illustrate the change of human decision-making over time.

Task 5: We introduce a risk-to-benefit factor and show how generational differences change our model. We also compare PI with PP and IP.

Task 6: To clarify the connection between different subgroups of people, the multi-dimensional clustering algorithm for friends (*mCAF*) is applied to the dataset *PIDATA*. By conducting experiments on the data from different groups as well as from the same group, we find that the relationship between data and value is not linear, but log-likelihood.

Task 7: We simulate the effect of massive data breach and predict the effect of PI loss and cascade event using our model. Based on our pricing system, we think agencies should compensate to individuals directly for data breaches.

In the end, we make sensitivity analysis and discuss the strengths as well as weaknesses of our model. Moreover, a policy memo is presented to the decision maker on the utility, results and recommendations based on our *PIPE* policy model.

Keywords: Private Information; Pricing Strategy; Dynamic System; Network Effect

PIPE: Estimate the Value of Private Information

Contents

1 Introduction

1.1 Problem Background

We are moving towards a "Web of the world" in which mobile communications, social technologies and sensors are connecting people, the Internet and the physical world into one interconnected network [1]. Vast quantities of data records are increasingly gathered by cheap and numerous information-sensing devices on personal information (PI) including but not limited to tweets, purchasing histories and health records. In 2016, roughly 16.1 zettabytes (10^{21} bytes) of data are being generated each day, and it is estimated that the figure will increase to 163 zettabytes by the year 2025 [2].

Mining and analyzing such data enables researchers to study, understand and even predict human behaviors on the individual, group and global level. Advanced data analytics methods that extract value from data have been in widespread use in insurance, marketing and many other industries [3]. For instance, methods combining big data with deep learning methods have shown superior performance in predicting traffic flows [4] and managing high-risk patients [5].

However, the massive collection, sharing and distribution of personal data are prone to certain risks concerning *information privacy*. As participation in social networking sites has dramatically increased in recent years, services such as *Wechat*, *Twitter*, and *Facebook* allow millions of individuals to create online profiles and share personal information with vast networks of friends-and, often, unknown numbers of strangers [6]. *Data breaches* also pose considerable threats to sensitive private information that involves personal health information (PHI), personally identifiable information (PII), trade secrets of corporations or intellectual properties [7].

It has been acknowledged that data providers can possibly be classified into subgroups according to their data's value distribution over multiple domains (e.g. finance, health). On the other hand, personal or community risks related to data privacy often arouse significant differences in peoples' privacy choices across such domains as well [8].

More and more intensive sharing are taking place nowadays, while the management and trading of private data are under loose control of the government and companies. Currently, millions of people are tricked into offering their data in exchange for little reward. However, the use of their data is far from efficient due to data isolation between enterprises. Moreover, some of these data are not even kept safe, and stolen data can possibly encourage illegal activities, such as fraud.

1.2 Our Work

To address this situation, we model private information that can be classified into several categories as a range of digital commodities that are constantly produced throughout a person's life, the value of which is determined by a joint strategy that takes into consideration potential losses caused by disclosure of personal information as well as social and commercial benefits to be exploited from that data. The actual price of such data fluctuates around its real value under the influence of supply and demand, the cumulative effect and many other factors.

In this paper, we introduce three feasible techniques. Firstly, we propose a vector-based representation for both data providers and data query requests that abstractly and quanti-

tatively describes features of private data, along with an corresponding value predictor that approximates the value function via correlation matrices. Secondly, with the introduction of a dynamic market system, we are able to further investigate the fluctuation in the real price influenced by both inner factors (e.g. supply and demand) and external causes (e.g. a sudden data bleach). Lastly, we develop an social network model to especially investigate the network effects of data sharing and the impact of social connection on data correlations with the multi-dimensional clustering algorithm *mCAF*.

The major contribution of this work is that we present a reliable price model for pervasive collection, sharing and trading of private information. In our experiments, we apply and test our model under diverse conditions, where it gives interesting and reasonable results which convinces us that the currency of private information should be kept under strict control under laws and regulations in order to maintain a healthy data economy.

2 Assumptions & Nomenclature

2.1 Assumptions

To better quantify the problem, our private information pricing model is based on several assumptions that hold true in most cases or is indisputably satisfiable under government regulation.

Assumption 1. *All kinds of private information can be classified into a fixed number of distinct* data categories *(e.g. demographics, family & health, etc.), the number is denoted by* m*.*

Assumption 2. *Personal data brings benefits to the society by contributing to researches that intends to study the social and financial behaviors. Profits made from fraud or harassment are not taken into consideration.*

Assumption 3. *For information security and many other concerns, all the gathered data are managed by a trusted third-party organization, which protects uses' data and helps sell them under owner's permission.*

Assumption 1 ensures that the number of parameters required to model private data is limited, thus it makes sense to represent PI with matrices. Assumption 2 guarantees that an universal understanding of data value exists, which forms the basis of our model. By assumption 3 large-scale management and regulation of data are made possible.

2.2 Nomenclature

In this paper we use the nomenclature in Table 1 to describe our model. Other symbols that are used only once will be described later.

3 PIPE: Mathematical Model for Private Information Price Estimation

In this section, we will discuss all details about our model, which is capable of establishing an accurate pricing system of personal data with the application of 1) a vector-based representation that distributes both benefits and risks of data from a certain subgroup over

m data categories; 2) a dynamic pricing strategy that determines the intrinsic value as well as market price of data; 3) a social network model that further improves the predicting accuracy by taking data correlation originated from social connections into account.

Table 1: Nomenclature

Symbol	Definition
m	Total number of data categories
c_i	The i^{th} category
I	Individual that produces data
X	Individual feature vector
$\sigma(\cdot)$	Individual feature extractor
q	PI query request
Y	Query feature vector
$\varphi(\cdot)$	Query feature extractor
C	Correlation matrix
v	Raw value of a person's data under a certain query
T	Sequence length of personal data
t	Freshness of private data
N	Quantity of data records
ω	Cumulative factor of data sequence
τ	Decay factor of history data
μ	Scale factor
v'	Amended value of a person's data under a certain query
d_i	Data size from person i
$\mathcal{I}$	Total information contained from the PI
Γ	Neighborhood
R	Region
Q	Types of Agencies
P_{agency}	Price concerning different agencies
G_{ij}	Interaction value between person i and j of social circle
O	Organizations
T_i	Tie strength
W_i^k	Weight summary of one measurement to one node
$Sim_{i,j}$	Similarity between two vertices
N	Threshold Neighbor
$\mathcal{M}$	Metric that evaluates the sensitivity of mCAF
$Eq(\cdot)$	Function that judges equality
l_i	Group label of vertex i

Our idea is that the intrinsic value of personal data comes from two aspects: the potential risk from information disclosure, and the social benefits brought about by data analytic. Such benefits and threats posed by personal data varies not only among different data categories, but also between diverse social subgroups. Another unnegligible factor that affects the value of data is the data quality demanded by corporations or institutes. For instance, a commercial dataset that requests detailed financial information should definitely be charged higher than a rough portrait only involving the overall income and tax bills of the same so-

cial group. There is nothing ambiguous that a variety of additional factors also have impacts on data value, including freshness, quantity and consistency, which are taken into account in our model as well.

Based on such assumptions we model private data as a commodity in continuous production, which is then fit into a market model where data owners can choose to pay for various levels of privacy protection, or to put their data on sale via a trusted third-party data manager. The overall supply is mainly determined by people's willing to share their data based on its benefit-risk ratio , while the market demand follows a gaussian distribution and can be affected by incidents such as data breaches. We further classify data agencies into three types based on their purchasing power, and develop a pricing strategy for the third-party data manager.

Since human data is highly linked and individual behaviors can be quite correlated with those whom they are socially, professionally, economically or demographically connected, we further consider the natural social network as a graph and cluster similar individuals on multiple dimensions based on both network structure and profile information. On the basis of such social clusters, we especially polish our pricing policy in consideration of similarities within clusters and distinctions between them.

The overview of our entire pricing framework and three major components of it are illustrated in Fig. 1.

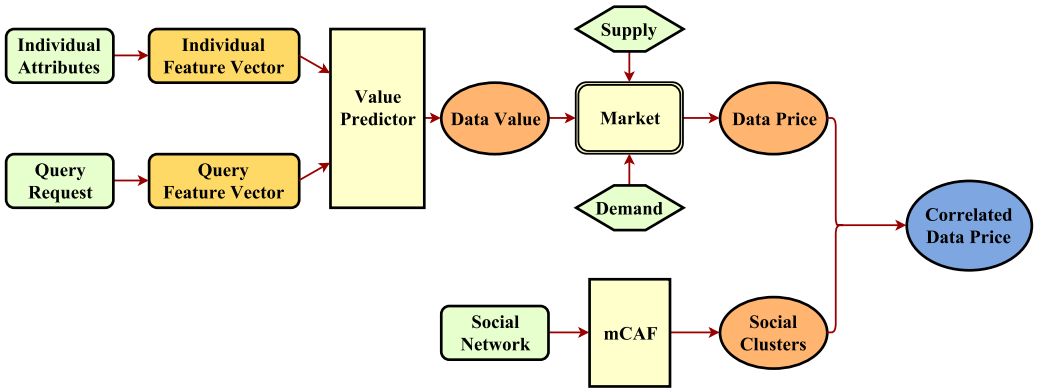


Figure 1: The schematic illustration of the entire model

3.1 Vector-based Representation for Individuals and Queries

3.1.1 Individual Feature Vectors

Based on the assumption in Sec. 2.1 that private data can be classified into m distinct categories, an individual's private data can thus be considered to consist of data records in multiple categories. In our model, we assume $m = 5$ and the categories include *demographics*, *family & health*, *property*, *activities* and *consumer* data.

The value of a person's entire data can then be split into m independent *category values*, the sum of which is equal to the original data value. Therefore, it makes sense to represent private information of individuals with an $m \times 1$ matrix, which is actually a vector:

$$\sigma(I) = X = \begin{bmatrix} X_1 & X_2 & \dots & X_m \end{bmatrix}, \tag{1}$$

where I is a data provider, element $X_i (1 \leq i \leq m)$ in vector X indicates the value of I's data in the i^{th} category, and function $\sigma(\cdot)$ extracts such category values based on life events

and individual attributes. We define vector X as person I's *feature vector* mainly because its elements reveals the essential value of I's data that is distributed over m categories.

The core part of Eqn. (1) is the function $\sigma(\cdot)$ that maps a person to the corresponding feature vector. The process is accomplished based on an analysis on the most influential factors of data value conducted by *Financial Times* [9]. The report points out that contrary to popular belief, the value of private information does not increase linearly with its amount. In fact, general information about a person, such as their age, gender and location is worth a mere \$0.0005 per person, or \$0.50 per 1,000 people. It is certain milestones in a person's life that prompt major changes in data values, such as becoming a new parent, moving homes, getting engaged, buying a car, or going through a divorce.

As is mentioned above, the value of data limited to a certain category is defined as the sum of potential risks and benefits incurred by it. On the basis of the Financial Times report, we develop a sophisticated model that implements the function $\sigma(\cdot)$. Here we briefly introduce its mechanism, the complete implementation can be referred to in Appx. A

A person can possess a number of attributes at the same time, such as *being engaged*, *owning a home* and *current job*, and some values of these attributes can possibly incur risks or benefits if known by a data company, for example, *being engaged = true* and *current job = government officer*. Our model includes a databases that stores the economical value vectors of certain attribute-value pairs, as is shown in Table 2, factors regarding to personal properties, health conditions and activities are considered more important and attached higher values than others.

Table 2: Some of the most significant value vectors in our

Attribute condition	Value vector / \$
Being a millionaire	$[0.116 \quad 0 \quad 0 \quad 0 \quad 0]$
Having a heart disease	$[0 \quad 0.260 \quad 0 \quad 0 \quad 0]$
Registered at a real estate agency	$[0 \quad 0 \quad 0.105 \quad 0 \quad 0]$
Interested in foreign travel	$[0 \quad 0 \quad 0 \quad 0.135 \quad 0]$
Holding a store loyalty card	$[0 \quad 0 \quad 0 \quad 0 \quad 0.136]$

By significance we mean the magnitude of vectors calculated by norm $||X||$. Note that all our vectors concentrate values in one dimension (one data category), by which we intend to reduce data correlations between categories, which will be reconsidered in Sec. 3.1.2.

Algorithm 1: Feature vector extractor

Input : An individual I, information value database $\mathcal{S}$.
Output: The feature vector $X = \sigma(P)$.

$X \leftarrow \begin{bmatrix}0 & 0 & \dots & 0\end{bmatrix}$;
for $attr \in \{P$*'s attributes*$\}$ **do**
  if $\langle attr, P[attr]\rangle \in \mathcal{S}.values$ **then**
    $X \leftarrow X + \mathcal{S}[\langle attr, P[attr]\rangle]$;
return X;

As is shown in Alg. 1, our algorithm first sets the initial feature vector as an all-zero vector and then checks all the attributes of that person to determine if some attribute-value

pairs can be found in our database. If a match is found, which means a certain value-creating condition is met, the value vector corresponding to that attribute-value pair will be added to the person's feature vector. In other words, a person's feature vector is the sum of all value vectors of the conditions satisfied by his personal attributes.

Based on Alg. 1, we are able to determine the typical individual feature vectors of various social subgroups, which will be discussed in Sec. 4.1.

3.1.2 Query Feature Vector & Correlation Matrix

Similar to *individual feature vectors* defined in Sec. 3.1.1, we define an $m \times 1$ *query feature vector* $Y = \varphi(q)$ that represents the query request q with m scalars, each stands for the intensity of private data requested in an data category. The intensity is measured by the amount as well as accuracy of requested data. For instance, if full information of a person's demographic characteristics and purchasing history is requested, the query feature vector should be $\begin{bmatrix}1 & 0 & 0 & 0 & 1\end{bmatrix}$.

However, with merely query vectors we are still not able to accurately calculate the value of private information, as data correlations tend to occur between different categories of data. To address this phenomenon, we introduce a *correlation matrix* C that takes connections between various categories of data into consideration, and define that the value of a piece of data record from individual I queried with feature vector Y as

$$v = \sigma(P)CY^T \tag{2}$$

Ideally, with no data correlations the correlation matrix $C = E = diag\left(\begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}\right)$. In order to estimate the intensity of data correlations, we fill the correlation matrix C in following manners:

$$C_{i,j} = |cov(c_i, c_j)| \tag{3}$$

where c_i and c_j are the i^{th} and j^{th} data category value. Our final correlation matrix C is computed based on feature vectors extracted from typical population subgroups, which will be discussed in detail in Sec. 4.1 and demonstrated in Fig. 6.

$$C = \begin{bmatrix} 1.0000 & 0.6463 & 0.8443 & 0.8231 & 0.2793 \\ 0.6463 & 1.0000 & 0.6767 & 0.2197 & 0.2226 \\ 0.8443 & 0.6767 & 1.0000 & 0.5403 & 0.1649 \\ 0.8231 & 0.2197 & 0.5403 & 1.0000 & 0.6916 \\ 0.2793 & 0.2226 & 0.1649 & 0.6916 & 1.0000 \end{bmatrix} \tag{4}$$

which suggests that the most magnificent data correlations exists between data categories of

- Demographics & Property
- Demographics & Activities
- Activities & Consumer

On the basis of Eqn. (2) and Eqn. (4), we are able to calculate the raw value of a specific data record and its query requests. However, there remain external factors that have strongly affect the real value of private data, among which the most significant one is time. It is widely acknowledged that data value decays with time. On the other hand, a consistent

data record sequence collected throughout a long time period should be attached additional value. Similarly, data scale affects the value of private data nonlinearly. Thus, an amendment is made with Eqn. (2) by introducing variations:

$$v' = e^{\omega T - \tau t} N^{\mu} v \tag{5}$$

where the dynamic element T denotes the sequence length ($/days$) of data, t stands for the freshness of private data (days since the data is generated), and N represents the number of data records. Parameters ω, τ and μ affects the real value in exponential and multinomial manners. In our estimation based on information rules [10], $\omega = 1.28 \times 10^{-3}, \tau = 9.50 \times 10^{-4}$ and $\mu = 1.05$.

3.2 Dynamic Market System & Pricing Strategy

3.2.1 PI Demand Model

The demand d_i for the private data is originate from the individual i (who themselves might posses the data or can generate the data). The availability of d_i to the demander j is captured by a matrix $\mathcal{M}$: the larger $\mathcal{M}_{i,j}$, the larger a fraction of i will be demanded by j. Specially, the demand that $j \in J^+$ will see from i is $d_i \cdot \frac{\mathcal{M}_{i,j}}{\sum_{\omega \in J^+} P_{i,\omega}}$. The matrix $\mathcal{M}$ will in practice depend on time latencies, as well as political and cultural issues. It need not be symmetric. For the purpose of the general model, we are agnostic to the derivation of $\mathcal{M}$. An individual $j \in J^+$ will incur a risk of r_j per unit of demand; This risk is the result of data breach/leakage, which will bring some potential trouble to the person or the related business. To encourage individuals to provide the PI, the buyer offers payments P_i to the person i. These payments are different from person to person, and can be derived from the domains (e.g. social media, financial transactions, and health/medical records) and some properties of that person.

Different people have different tradeoffs between the price and risk of PI. We model this fact by assuming that each person j has a trade off factor λ_j that describes the risk-to-benefit ratio of PI. We use Benefit-Divide-Risk Analysis (BDRA) to calculate λ_j. It is defined in Eqn. (6), where $s_{benefit}$ and s_{risk} are benefit score and risk score. Benefit score is the data value we have calculated. Risk score is calculated in Table 3.

$$\lambda_j = \frac{s_{benefit}}{s_{benefit} + s_{risk}}, \lambda_j \in [0, 1] \tag{6}$$

Table 3: Calculation of risk score (1,2,3,4,5 stand for level of risk)

Criterion	1	2	3	4	5
Financial Risk	0.05	0.062	0.074	0.09	0.128
Health Risk	0.06	0.09	0.11	0.14	0.24
Family Risk	0.05	0.062	0.088	0.15	0.22
Social Risk	0.03	0.062	0.1	0.13	0.17

Thus, the *demanding* of an individual $j \in J^+$ is

$$D(j) = \lambda_j P_j - r_j \sum_i \frac{d_i \mathcal{M}_{i,j}}{\sum_{\omega \in J^+} \mathcal{M}_{i,\omega}}, \tag{7}$$

Asymmetric Information. This part illustrates that purchasers possess a substantial amount of information about data providers. These facts suggest that the restrictions such as policy and culture - which make it difficult for data providers to adjust prices when they provide private information to purchasers over time - have different price effects across different sellers and buyers depending on their privately revealed types, cultural and political issues. Incorporating such factors in our model therefore become important in anticipation of using the model to study the PI problem. Formally, we estimate these effects in the following equation 8,

$$Default_{i,t} = \alpha_{j(i)} + \alpha_t + \sum_{n=1}^{5} \beta_n 1_{\psi_i, t=n} + \epsilon_{it} \tag{8}$$

Here the dependent variable is an indicator for any instance of default by purchaser i after period t, and the key coefficients β_n capture differences in default rate across different private information, which denoted by ψ. Meanwhile the fixed effects for purchaser i and time t help ensure that these risk comparisons are made within otherwise observably similar purchasers.

Model Exposition This section presents the PI model. The backbone of the demand model is a finite mixture of agency types, each of whom has demand over PI providers.

Our model denotes type by θ. We specify several parameters to be estimated for each type. First, each type enjoys a flow utility $d_{j\theta}$ from buying from person j and a time utility $n_{j\theta}$ from transacting with person j; meanwhile the utility is normalized to zero. Additionally, in order to capture the adjustment cost, each type pays a agency cost $s_{j\theta}$ for refer to potential related PI with person j. The parameters $\{d_{j\theta}, n_{j\theta}, s_{j\theta}\}_{(\theta,j)\in\Theta\times J}$ are the key demand parameters to be estimated in the model, along with a probability distribution μ_θ over types.

Integrating over taste shocks ϵ for each choice yields the standard Bellman equation for continuation values V, which is shown in Eqn. (9).

$$V(\theta, j, k) = \log\left(\sum_{j',k'} \exp(v(j', k'|j, k, \theta))\right), \tag{9}$$

where the lower-case v term denotes total expected payoffs. The value of v depends on data buyers' past-period and current-period choices. The expectation $\mathbb{E}_\theta$ can be decomposed as Eqn. (10),

$$\mathbb{E}_\theta[V(\theta', j, b)] = (1 - \delta(\theta))T_{\theta\theta'}(\theta)V(\theta', j, b) + \delta(\theta)T_{\theta\theta'}(\theta)V(\theta'', 0, 0). \tag{10}$$

With the establishment of Eqn. (5), our model further takes time variations and scale effects as dynamic elements into consideration to estimate the worth of personal data over time.

3.2.2 Buyer-Seller Relationship Influence to Price

As Fig. 2 shows, internet advertising revenue has grown strongly over the last ten years. In 2013 it hit $42.8 billion in the US. Internet giants such as Google and Facebook have business models underlined by the use of personal data, but most people would have trouble knowing who exactly has access to the data trail they are generating across the internet [11]. A recent study by JPMorgan Chase [12] found that each unique user is worth approximately $4 to Facebook and $24 to Google.

Besides commercial corporations, there are also other agencies who purchase PI. Mozilla collect data about users to better personalize their experiences with their open source products such as Firefox, Thunderbird. The information they gather through analytics can be used to make their product easier to use. They also use cookies (small data files placed in browsers) to remember language preferences. Center for Disease Control utilizes the data shared to trace the spread of disease in order to prevent further outbreak.

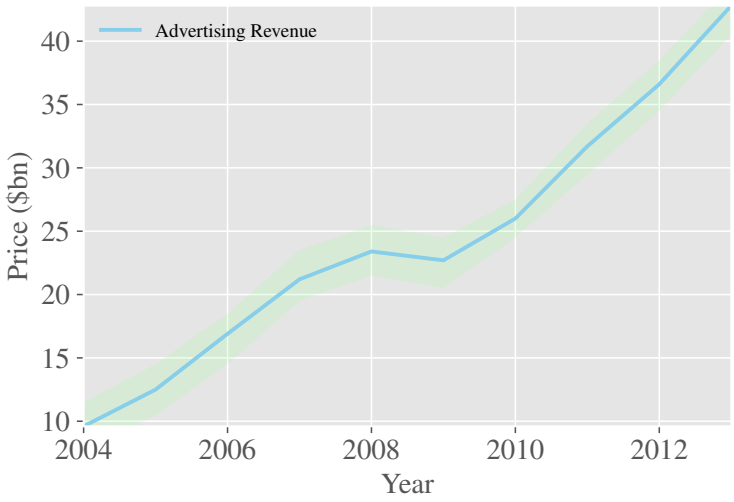


Figure 2: US Internet Advertising Revenue by Year, the shadow represents statistical uncertainty and variance.

There are 3 types of agencies in our model and their purchasing power is shown as Table 4. The price estimation system concerning with different agencies who purchased the PI is illustrated as Eqn. (11).

$$P_{agency} = \sum_{j=1}^{M} \tau_j \mathcal{Q}_i, \tag{11}$$

where τ_j is the control level of individual j to sell his/her own data, there are totally M individuals in a group/nation, $\mathcal{Q}_i$ can be $\mathcal{Q}_1, \mathcal{Q}_2$ or $\mathcal{Q}_3$ which represents the purchasing power of different types of agencies.

3.3 *mCAF*: a Multi-dimensional Clustering Algorithm for Friends of Social Network Services

The multi-dimensional clustering algorithm for friends (*mCAF*) is adopted by us to perform multi-dimensional clustering. Multi-dimensional clustering algorithms on social networks are progressively gaining popularity due to the information and insights produced using large-scale social data. [13] describes the user's opinions, comments, and likes in social media have significant relationships with the popularity of that post. Multi-dimensional cluster analysis is a strategy for identifying different Facebook users' fan groups and provides insights to prompt further research analytics [14]. Both network structures and profile information should be taken into consideration while analyzing a user's clusters on social networks [6, 15].

Table 4: Purchasing power of three types of agencies

Denotation	Types of Agencies
$\mathcal{Q}_1$	Commercial Corporation, e.g. Google, Facebook, Microsoft, etc.
$\mathcal{Q}_2$	Non-Profit Organization (NPO), e.g. Mozilla, GNU, WWF, etc.
$\mathcal{Q}_3$	Government Department, e.g. NSA, Department of Energy, etc.

In this study, we used the Facebook Graph API to retrieve information of 600 users. First, we define the measurements of clustering.

Social Circles. A social circle is a group of people who have the same interests or join the same activity. We define M_{ij} as the number of mutual friends of user i and j and G_{ij} as the interaction value. We quantify the subject's interactions within the community to obtain S_{ij} with Eqn. (12) and then normalize the result as Eqn. (13)

$$MG_{ij} = M_{ij} + G_{ij}, \mathbb{MG} = \{MG_{xy} \mid x, y \in 1, 2, 3, \ldots, n\} \tag{12}$$

$$S_{ij} = \frac{MG_{ij}}{\max(MG)}, \mathbb{S} = \{S_{xy} \mid x, y \in 1, 2, 3, \ldots, n\} \tag{13}$$

Regions. We determine the region of the users and then calculate the distances between them and store them as a dataset described by $\{D_1, D_2, D_3, D_4\}$. Take the calculation of distance between A and B as an example, D_1 represents the distance between the hometowns of A and B; D_2 represents the distance between the current residence of A and B; D_3 represents the distance between A's hometown and B's current residence; D_4 represents the distance between A's current residence and B's hometown. The calculation of R_{ij} is shown as Eqn. (14).

$$R_{ij} = \alpha \times D_1 + \beta \times D_2 + \gamma \times D_3 + \delta \times D_4 \tag{14}$$

Organizations. If two individuals attended the same school or worked in the same company, the organizations measurement O_{ij} is set equal to 1 since they have a connection. If no connection is present, the O_{ij} is set equal to 0.

Tie strength. We retrieve related information and use the method described in [16] to calculate the tie strength as T_j, which indicates the tie strength between a user and his jth friend.

mCAF maps a user's friends into un-directed, weighted graphs. We define the entire graph as $G = \{V, E\}$, in which V is the set of vertices and E is the set of edges, defined as $\left\{E_{i,j}(e^k_{i,j})\right\}$, which represents a connection if a value $e^k_{i,j}$ is greater than zero between nodes i and j under measurement k.

Definition of vertex structure. Let vertex $i \in V$, where the structure of i is defined by its neighborhood denoted by $\Gamma(i)$ in Eqn. (15)

$$\Gamma(i) = \{j \mid j \in V \wedge E_{i,j} \in E\} \tag{15}$$

Definition of the weight summary of one measurement to one node Eqn. (16) defines the summary values of measurements from vertex j, which is connected to i:

$$W^k_i = \sum_{j=1}^{j=|V|} (e^k_{i,j}), where j \in \Gamma(i) \tag{16}$$

Definition of the weight summary of one measurement to two nodes Let vertex $m \in V$, and let edges from (i, m) and (j, m) exist. Eqn. (17) defines the summary values of measurements from vertex m, which is connected to i and j:

$$T_{i,j}^{k} = \sum_{m=1}^{m=|V|} (e_{i,m}^{k} + e_{j,m}^{k}) \; where \; m \; \in \Gamma(i) \; and \; m \in \Gamma(j) \tag{17}$$

Algorithm 2: Multi-dimensional clustering algorithm(*mCAF*)

Input : $G = \{V, E\}, \{\epsilon^k\}, \{\mu^k\}$
Output: Clustering result

$count^1 = 0; count^2 = 0; count^3 = 0; k^1 = 0; k^2 = 0; k^3 = 0;$
foreach *vertex* $i \in V$ **do**
  for *each vertex* $j \in \Gamma(i)$ **do**
    if ($S_{i,j}^k \times \mu^k$ *has max value where* $k = 1 \sim 3$ *and* $j \in N_{\epsilon^k}(i)$ **then**
      $count^k = count^k + 1;$
  Set($|count^1|$, $|count^2|$, $|count^3|$) as (k_1, k_2, k_3);
  if k^a *has only one max value* **then**
    $k_{max}(i) = a$;
  else
    label i as uncertain V;
foreach *unstagged vertex* $p \in V$ **do**
  if $q \in N_{\epsilon^{k_{max(p)}}}(p)$ **then**
    if *first* q **then**
      generate new clusterID;
      insert q into queue Q;
    while $Q \neq 0$ **do**
      q=dequeue(Q);
      foreach $r \in N_{\epsilon^{k_{max(p)}}}(q)$ **do**
        if *(*r *is untagged)* **then**
          insert r into queue Q;
          assign current clusterID to r;
      remove q from Q;

Definition of structure similarity She structure similarity of two vertices i and j is defined as Eqn. (18):

$$Sim_{i,j} = \{S_{i,j}^1, S_{i,j}^2, S_{i,j}^3\} = \frac{\{T_{i,j}^1, T_{i,j}^2, T_{i,j}^3\}}{\sqrt{W_i^1 \cdot W_j^1 + W_i^2 \cdot W_j^2 + W_i^3 \cdot W_j^3}} \tag{18}$$

Definition of the threshold neighbor If two nodes can be clustered together based on measurement k, their structure similarity value $S_{i,j}^k$ must be greater than the preset threshold ϵ^k to filter out noise. Eqn (19) defines neighbors with qualified similarity structure values. The parameter ϵ^k could be estimated via training.

$$N_{\epsilon^k}(i) = \{j | j \in \Gamma(i) \wedge S_{i,j}^k \leq \epsilon^k\} \ where\ k = 1\ to\ 3 \tag{19}$$

The complete mCAF algorithm is described in Algorithm 2.

The clustering result of *mCAF* is shown in Fig. 3. The social network of several individuals are clustered into 8 subgroups. The visualization of network shows the correlations between different people.

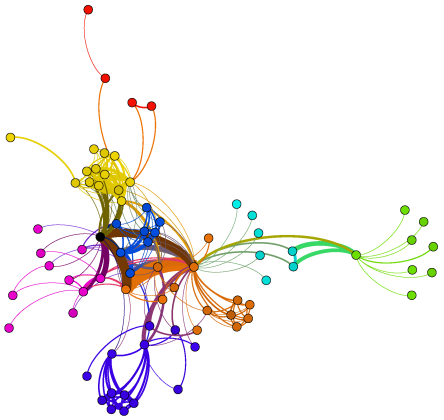


Figure 3: *mCAF* Clustering Result of Social Network

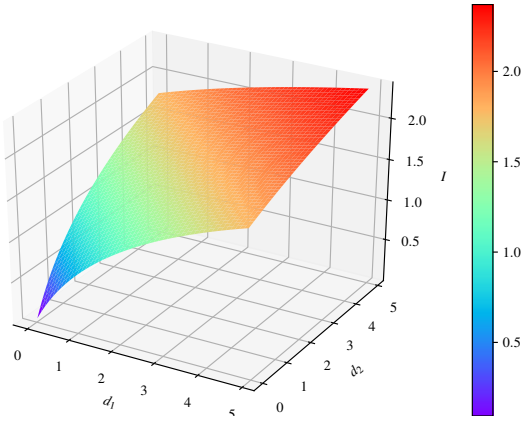


Figure 4: Relationship between data size d_i and information $\mathcal{I}$

To better evaluate the network effects of data sharing, we investigate the relationship between two individuals who are highly linked and discover the relationship between the data size from one person and the information it can provide. Define d_i is the data size from person i and $\mathcal{I} = f(d) \in [0, 1]$ represents the information the data on one person can provide. There are totally M individuals. Eqn. (20) represents the function between d and $\mathcal{I}$ and Fig. 4 illustrates the function in the circumstance of $i = 2$.

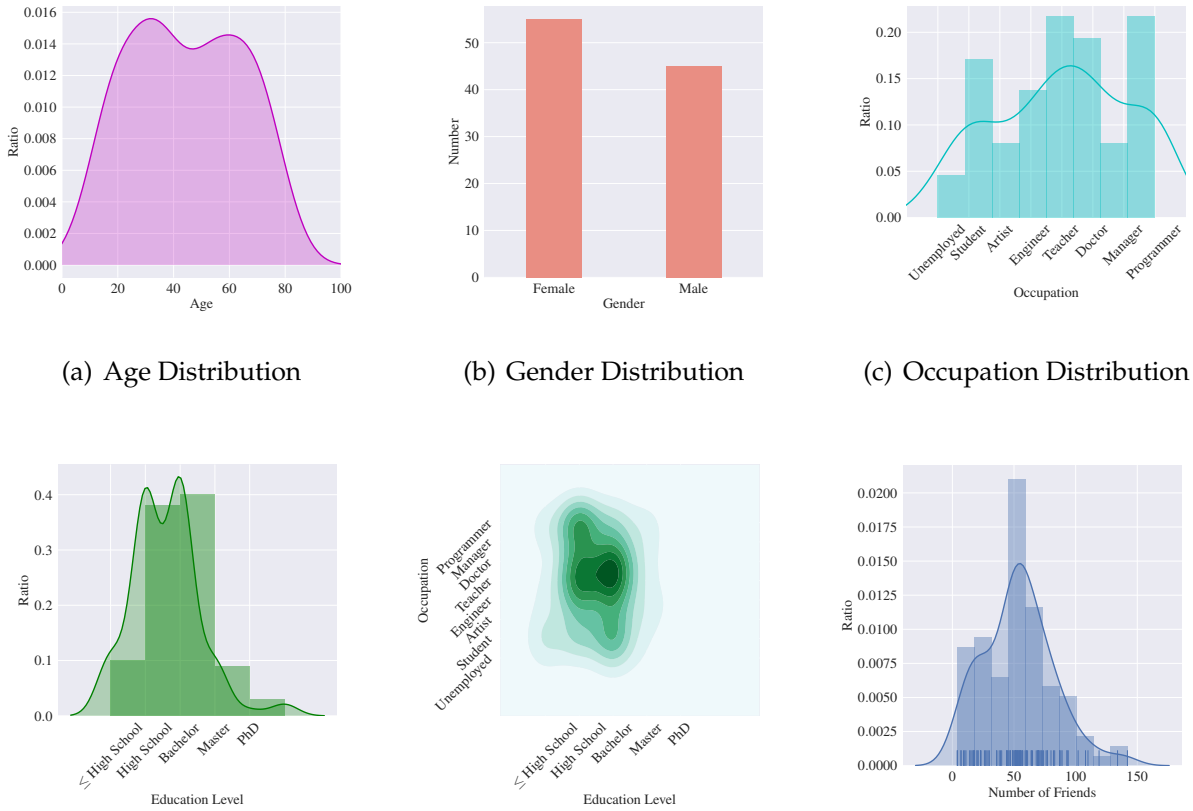
$$\mathcal{I} = \log(1 + \sum_{i=1}^{M} d_i) \tag{20}$$

4 Experimental Results

4.1 Task 1: Price Point for Protecting One's Privacy and PI in Various Applications

In order to accurately model risk to account for both 1) characteristics of the individuals, and 2) characteristics of the specific domain of information, we introduce the concept of individual and query feature vectors discussed in Sec. 3.1.

After surveying on several datasets [17, 18, 19] and recent methods, we collect our dataset *PIDATA* with an API provided by Facebook [20]. The usage of this dataset observes the *Platform Policies*, *Data Use Policy*, *Statement of Rights and Responsibilities*. Corresponding statistical information is illustrated as Fig. 5. Fig 5(a) is age distribution. Fig 5(b) is gender distribution. Fig 5(d) is education distribution. Fig 5(c) is occupation distribution. Fig 5(e) is education occupation distribution. Fig 5(f) is friend distribution.



(a) Age Distribution (b) Gender Distribution (c) Occupation Distribution

(d) Education Distribution (e) Education - Occupation Joint Distribution (f) Number of friends

Figure 5: Statistic information of *PIDATA*

Fig. 6 shows the distribution of private data over different data categories (feature vectors) of 6 selected subgroups. It is clear that the value of person's data is partially related his age and social status. Another interesting observation from Fig. 6 is that high covariance exists between certain data categories, which will be dealt with in Sec. 3.1.2.

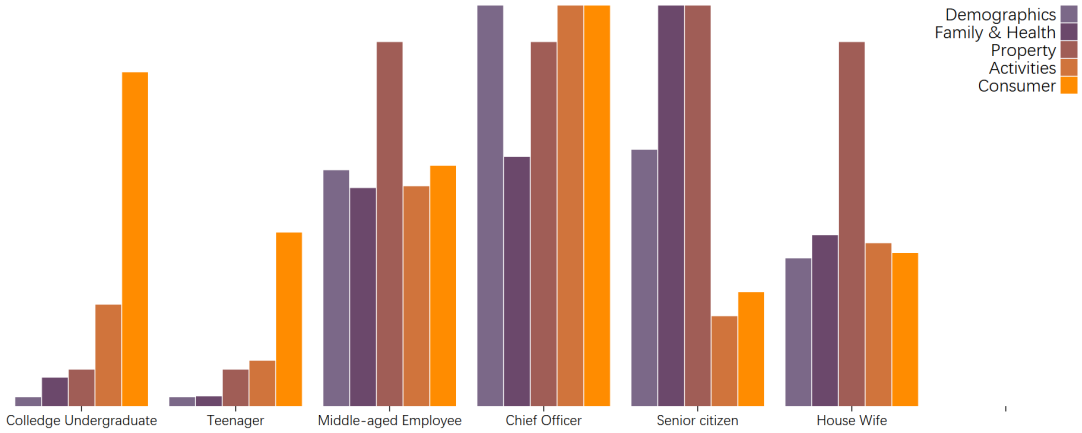


Figure 6: Individual feature vectors of some typical subgroups.

To develop a price point for PI protection, we exam the true value of a person's full data and its components. Fig. 7 shows how different categories of private data contributes to the value of a person's PI.

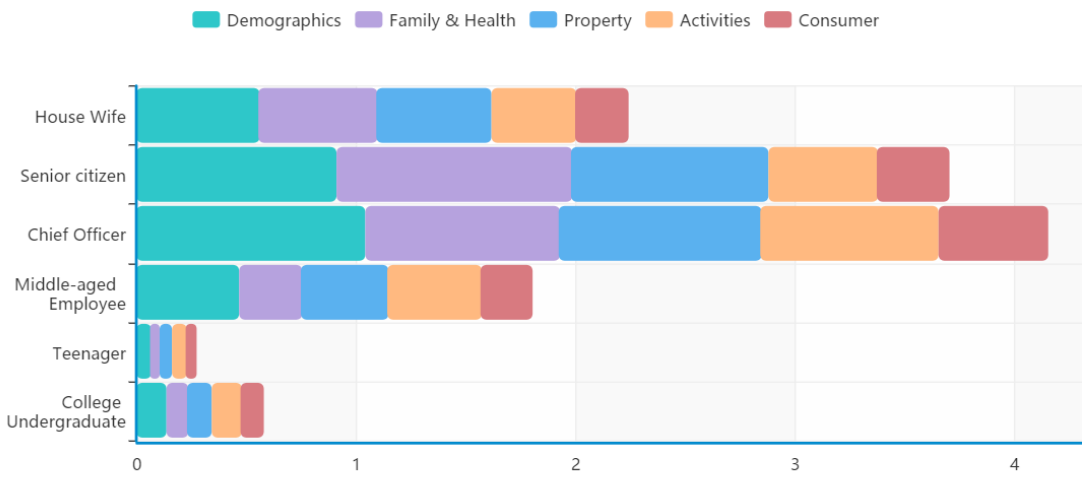


Figure 7: Value components of different subgroups's PI.

Based on the computed value of personal information, we are able to establish a price point for protecting it by treating data protection as an special insurance. Although the trade value of data can is just a few dollars, cost for each stolen data record can be as high as $141 - roughly 70-fold of its trade value, and the likelihood of a recurring material data breach over the next two years is estimated as 27.7%, according to a data breach study by IBM [21]. Therefore, to firmly protect a certain data record for 1 year, one must invest 10 times of its trade value calculated from Eqn. (5), which leads to the deduction that it might be better to share personal data and make profits from it if the risk is low.

4.2 Task 2: Pricing Structure of PI

With the introduction of feature vectors, correlation matrix and the amendment formula (Eqn. (5)), we can simply determine how much a person's information is worth given a query on a specific domain. Based on a thorough survey on 77 adults, we are able to estimate the corresponding query feature vector of several query domains. The results are listed in Table 5.

Based on query feature vectors in Table 5 and Eqn. (2), values of private information queried by different domains are illustrated in Fig. 8.

From results shown in Fig. 8, we can establish a pricing structure correspondingly. The price is explicitly calculated in view of different basic elements of data via individual feature vectors defined in Sec. 3.1.1. As for cost of privacy across various domains, it can be

Table 5: Comparison between query feature vectors from various domains

Domain	Average query feature vector
Social media	[0.93 0.68 0.13 0.96 0.37]
Financial transactions	[0.82 0.53 0.94 0.24 0.98]
Health / media records	[0.26 0.99 0.08 0.11 0.10]
Search histories	[0.66 0.58 0.13 0.74 0.84]
Location info	[0.24 0.05 0.05 0.35 0.12]

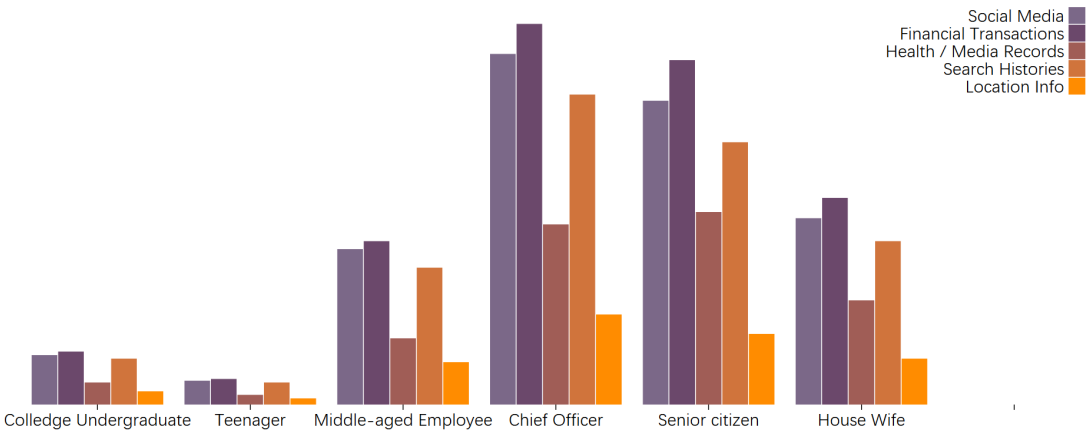


Figure 8: Information value in various domains of some typical subgroups.

calculated in similar manners as Sec. **??**, where protecting cost and trade value of data are in direct proportion. If the risk of data disclosure is negligible, spending considerable money on data protection would be an unnecessary cost, and people might prefer to have their data unprotected in order to save budgets.

4.3 Task 3: Supply and Demand

People become more clear about which agencies had purchased their PI, how much their PI was worth and how PI was being used.

Based on the above model, we evaluate the influence of the control level to the price of PI. With data becoming a commodity, we find that:

- It is appropriate to consider forces of supply and demand for PI. Commercial Corporations $\mathcal{Q}_1$ have higher demand for PI, which makes it possible for them to provide higher offer compared with the other two types of agencies.
- If people have control to sell to their data, which means τ_j varies with different individuals, the price P_{agency} increases with the τ_j.

4.4 Task 4: Assumptions and Constraints - Political/Cultural Issues

The assumptions and constraints of our model, which is also the political and cultural issues of the United States, European Union and other countries is listed as below. Suppose our model is proposed under the circumstance which is in compliance with local government regulations and cultural issues. Firstly, we explain the terminology used in the model.

4.4.1 Explanation of Terminology

- **Agency:** any corporations, organizations, Executive department, military department, Government corporation, Government controlled corporation, or other establishment in the executive branch of the Federal Government.
- **Individual:** a citizen of the United States or an alien lawfully admitted for permanent residence.

- **Private Information:** any item, collection, or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a photograph.

4.4.2 Political Issues and Cultural Issues

The United States: *Privacy Act* [22]

The *Overview of the Privacy Act of 1974, 2015 Edition* is prepared by the Department of Justice's Office of Privacy and Civil Liberties (OPCL). Tracking the provisions of the Act itself, the Overview provides reference to and legal analysis of court decisions interpreting the Act's provisions.

The purpose of the *Privacy Act* is to balance the government's need to maintain information about individuals with the rights of individuals to be protected against unwarranted invasions of their privacy stemming from federal agencies' collection, maintenance, use, and disclosure of personal information about them. More details are in the Appx. B.

European Union: *General Data Protection Regulation (GDPR)* [23]

The GDPR aims primarily to give control back to citizens and residents over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU. The regulation was adopted on 27 April 2016. It becomes enforceable from 25 May 2018 after a two-year transition period.

Data breaches. Under the GDPR, the Data Controller will be under a legal obligation to notify the Supervisory Authority (SA) without undue delay. The reporting of a data breach is not subject to any *de minimis* standard and must be reported to the Supervisory Authority within 72 hours after having become aware of the data breach.

Citizen Control of Personal Data. Under the GDPR, organizations are encouraged to give back control of personal data to the individual, or citizen.

Canada: *Personal Information Protection and Electronic Documents Act*

The Personal Information Protection and Electronic Documents Act (PIPEDA or the PIPED Act) is a Canadian law relating to data privacy. It governs how private sector organizations collect, use and disclose personal information in the course of commercial business. In addition, the Act contains various provisions to facilitate the use of electronic documents.

The law gives individuals the right to

- know why an organization collects, uses or discloses their personal information;
- know who in the organization is responsible for protecting their personal information;
- expect an organization to protect their personal information by taking appropriate security measures;
- obtain access to their personal information and ask for corrections if necessary; and

Japan: *Act on Protection of Personal Information (APPI)*

APPI reflects the Japanese socio-cultural characteristics for personal information protection. Personal information leakage cases and social responses in Japan reflect three Japanese socio-cultural characteristics: Uchi/Soto awareness, insular collectivism and Hon'ne/Tatemae tradition. An effective law protecting personal information in Japan's cultural environment cannot be made simply by copying the privacy protection laws in western nations. Instead, legal protection of personal information should be drafted that reflects and takes into account these socio-cultural characteristics [24].

4.4.3 Price Regulations

Generally, the data value of people varies from different regions. It's common to see quite a lot of variation by location. Fig. 9 is the statistical result, which is a comparison of the total number of a person's data value for countries around the world (ones without enough data are left gray) [25]:

Based on the model and the political/cultural issues, we can draw the conclusion that information privacy should be made a basic human right when thinking about policy recommendations.

4.5 Task 5: Generation Difference

In the perspective of risk-to-benefit ratio of PI and data privacy, there are generational differences. For example, the risk-to-benefit ratio is different between old people and young people when their health record is leaked. For old people, it is usually much higher than that of young people.

As generation changes, the input individual attributes change and data value depends on the generation correspondingly. As for the risk-to-benefit factor we defined in Eqn. (6), it describes the risk-to-benefit ratio of PI. The factor will change since the benefit score and risk score will change. Considering all the effects that generation change have, our final data price will be affected in the process above. PI (private information) is different from PP (private personal property) and IP (intellectual property). PI is an abstract concept. It is non-entity. However PP often refers to the property of people. It physically exists. IP is

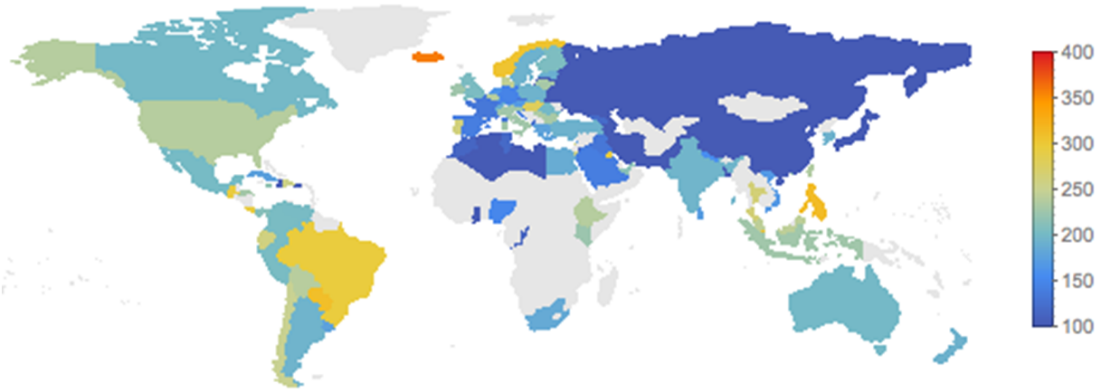


Figure 9: Data Value by locations: comparisons of the total number of a person's data value for countries around the world (ones without enough data are left gray)

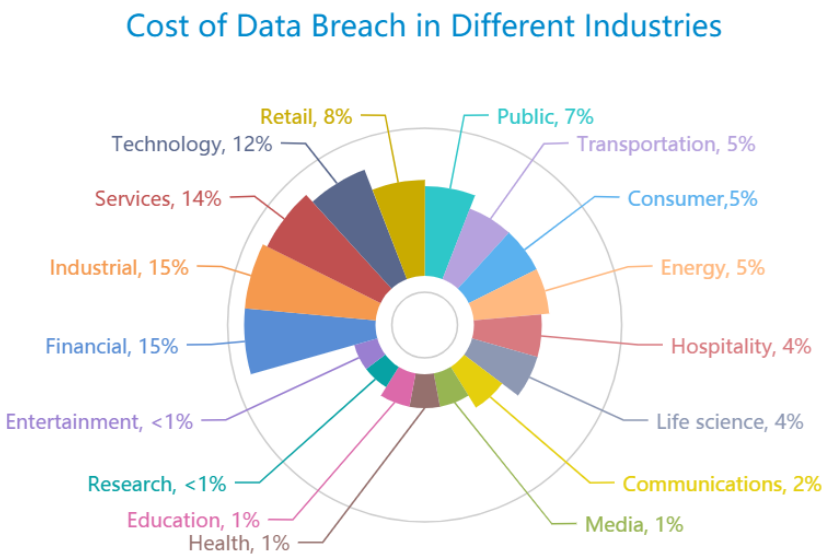


Figure 10: Cost of Data Breach in Different Industries

the property of human intellect including copyrights and patents. PI is only known by the owner but IP is not only known by the owner. People all know the owner of the IP.

PI is also similar with PP and IP to some extent. PI and PP are both private while PI and PP are both abstract.

4.6 Task 6: *mCAF*: a Multi-dimensional Clustering Algorithm for Friends of Social Network Services

Based on the *mCAF* model we proposed previously, we have the following conclusions:

- Network effects of data sharing sharing do effect the price system for individuals, subgroups, and entire communities and nations.
- It is the responsibility of the communities to protect citizen's PI if communities have shared privacy risks.

4.7 Task 7: Data Breach Effect

Data breach, especially massive data breach where millions of prople's PI are stolen will affect the privacy a lot. Taking TJX data breach as example, it involves more than 100 million records and causes 118 million dollars loss to cover the loss and potential liabilities. That does not include the loss in reputation of brand and other indirect cost. From the research of IBM [21], we get cost of data breach for different industries (Fig. 10) and countries (Fig. 11). We can see that data breach widely exists and PI loss is a factor that we should not neglect.

The PI loss and cascade event caused by massive data breach will impact the price point. In our model, we have considered such effect. Eqn. (7) shows the trade off between the price and data breach effect. When massive data breach happens, the trade off factor λ_j will be smaller and thus $\lambda_j p_j$ will be smaller. r_j depends on the level of data breach. For massive data breach and the corresponding cascade event, it would be higher. We can see that the

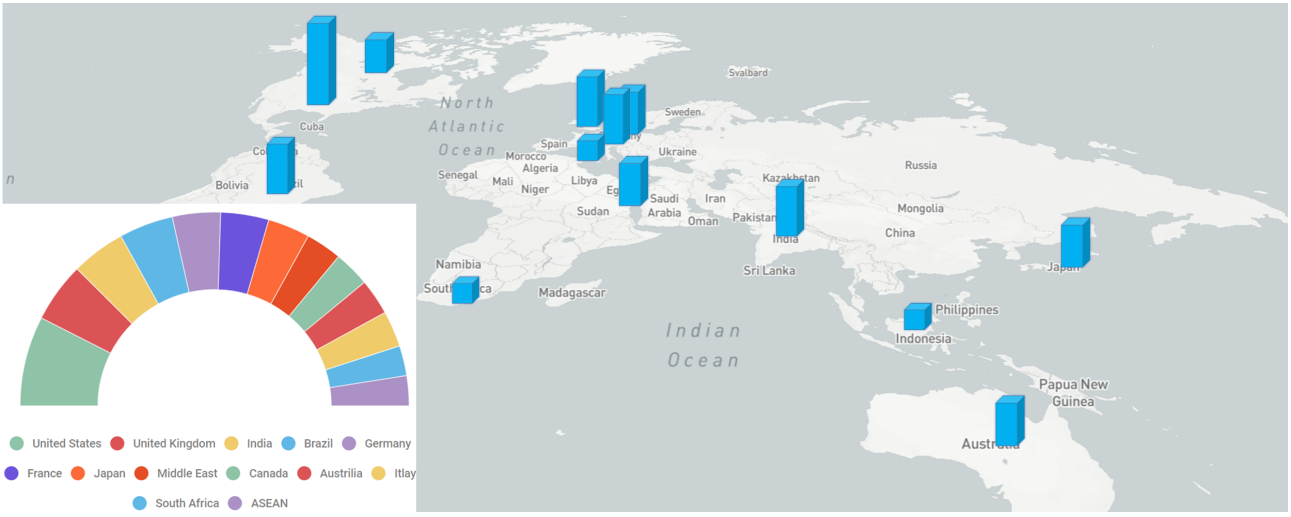


Figure 11: Cost of Data Breach in Different Regions

price will be lower. It is reasonable since data breach will violate the assumption that all the gathered data are managed by a trusted third-party organization, which protects users' data and helps sell them under owner's permission. People's data will be sold without permission or even be used as ransom. Some PI buyers will buy the breach data in an illegal way instead of buying data from people or trusted third-party organization. Demand will decrease and the price will be lower.

Agencies that breach the data should be responsible for the data breach and pay the individuals directly. As shown above, price of data will be much lower after massive data breach happens. The agencies should be responsible for the PI loss even if they don't intend to breach the data.

5 Sensitivity Analysis

In this part we will do sensitivity analysis on our model. The sensitivity analysis show that our model is generalized and performs stably under different conditions, by which we are convinced that our model is able to solve the problem successfully.

5.1 Demand Model

This part shows the PI purchasers are sensitive to price and illustrates how it is possible to identify heterogeneous price sensitivities in the data. This heterogeneity will play a key role when we later use the model to study the equilibrium effects of the PI's price restrictions, because this heterogeneity affects which types of purchasers change their purchasing behavior in response to different relative price changes.

The event-time-specific estimates is shown as Eqn. (21):

$$\log Q_{jt} = \alpha_j + \alpha_t + \beta_j t + \alpha_{A,t} + \epsilon_{jt}, \tag{21}$$

where Q_{jt} denotes retention rates among existing buyers for data provider j at time t. The first two α terms in this equation implement a standard difference-in-difference design, while the $\alpha_{A,t}$ terms capture differences between data provider A and other. For sake of

presentation, the β term is included to account for different time trends among the included data providers. When estimating the demand side of the model, we use price variation to estimate heterogeneous price sensitivities across different purchaser types.

5.2 *mCAF* Model

We adopt the multi-dimensional clustering algorithm for friends (*mCAF*) to perform identify social clusters. mCAF algorithm randomizes the center points initially. We run mCAF several times on our dataset *PIDATA* and analyze the result.

We propose a metric to evaluate the sensitivity of mCAF in Eqn. (22), where $Eq(\cdot)$ is defined in Eqn. (23)

$$\delta = \frac{\sum_{i \in V} Eq(l_i, l_{0_i})}{\sum_{i \in V} 1} \tag{22}$$

$$Eq(x, y) = \begin{cases} 1 & \text{if } x \neq \text{ y} \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Table 6 shows the results from 8 experiments, from which we can conclude that the δ is very small in all experiments. It proofs that mCAF model perfoms well and is robust under different situations.

Table 6: Result of Sensitivity Analysis on mCAF

$ExpNo$	1	2	3	4	5	6	7	8
δ	0.025	0.013	0.025	0.013	0.051	0.013	0.013	0.025

6 Conclusions and Future Work

We develop a complete pricing system that accurately estimates the intrinsic value as well as market price of a certain individual's data under a specific query request. Out method consists of three core components: a value calculator that maps a query and an individual to value of that data, a dynamic market system to further compute its market price, and a social cluster model to estimate the network effect on data price. In the environments and sensitivity analysis of our model, we find that it is accurate and generalized enough to be adopted in a wide range of domains.

One limitation of our current approach is that our model requires a large dataset to precisely estimate the correlation matrix as well as many other parameters used in its mechanism. A promising future direction is to combine out methods with better estimation algorithms, e.g. Analytic Hierarchy Process (AHP), to reduce the variance caused by insufficient data.

In future work, our goal is to introduce non-linear data value predictors, e.g. neural networks, to characterize a more complicated relationship between individuals and data. Additionally, we plan to extend the current dataset in both scale and coverage to cater to the needs of deep learning algorithms.

References

[1] M. Davis, R. Martinez, and C. Kalaboukis, "Rethinking personal information - workshop pre-read." Invention Arts and World Economic Forum, 2010.

[2] R. David, G. John, and R. John, "Data age 2025: The evolution of data to life-critical," seagate.com. Framingham, MA, US: International Data Corporation, Tech. Rep., 04 2017.

[3] "Big data," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Big_data

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[5] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[6] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *ACM Workshop on Privacy in the Electronic Society*, 2005, pp. 71–80.

[7] "Data breach," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Data_breach

[8] B. Debatin, J. P. Lovejoy, A.-K. Horn, and B. N. Hughes, "Facebook and online privacy: Attitudes, behaviors, and unintended consequences," *Journal of Computer-Mediated Communication*, vol. 15, no. 1, pp. 83–108, 2009.

[9] E. Steel, C. Locke, E. Cadman, and B. Freese, "How much is your personal data worth?" https://ig.ft.com/how-much-is-your-personal-data-worth, Financial Times, 2013.

[10] C. Shapiro and H. R. Varian, *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press, 1998.

[11] J. Detemple and M. Rindisbacher, *The Private Information Price of Risk*. Palgrave Macmillan UK, 2016.

[12] J. Brustein, "Start-ups seek to help users put a price on their personal data," *The New York Times*, vol. 12, no. 3, 2012.

[13] H. Khobzi and B. Teimourpour, "How significant are users' opinions in social media?" *International Journal of Accounting & Information Management*, vol. 22, no. 4, pp. 254–272, 2014.

[14] E. Wallace, I. Buil, L. de Chernatony, and M. Hogan, "Who "likes" you and why? a typology of facebook fans: From "fan"-atics and self-expressives to utilitarians and authentics," *Journal of Advertising Research*, vol. 54, no. 1, pp. 92–109, 2014.

[15] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 1, pp. 1–28, 2014.

[16] T. T-H, C. H-T, C. Y-J, H. Y-H, L. D-H, K. C-C, and Y. T-Y, "TreeIt: an application to create, maintain, and enhance online social connections," in *Networking and Electronic Commerce Conference (NAEC)*. NAEC2014, 2014.

[17] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM Review*, vol. 53, no. 3, pp. 526–543, 2011.

[18] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS One*, vol. 8, no. 9, p. e73791, 2013.

[19] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta*, 2010.

[20] W. Graham, *Facebook API developers guide*. Infobase Publishing, 2008.

[21] P. Allor, "Cost of data breach study," https://www.ibm.com/security/data-breach, 2017.

[22] J. T. O'Reilly, "The privacy act of 1974." *Censorship*, vol. 61, no. 2, p. 7, 1975.

[23] G. D. P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46," *Official Journal of the European Union (OJ)*, vol. 59, pp. 1–88, 2016.

[24] Y. Orito and K. Murata, "Socio-cultural analysis of personal information leakage in japan," *Journal of Information, Communication and Ethics in Society*, vol. 6, no. 2, pp. 161–171, 2008.

[25] V. Gkatzelis, C. Aperjis, and B. A. Huberman, "Pricing private data," *Electronic Markets*, vol. 25, no. 2, pp. 109–123, 2015.

A Implementation of Function $\sigma(\cdot)$

We format our $\langle attribute, value \rangle$ database in the form of a reader-friendly questionnaire consisting of five parts, each corresponding to a data category as value vectors in our database concentrate values in one dimension. The five scores generated by the following questionnaire make up for the five elements in the individual feature vector of the tester.

A.1 Demographics

Has the following information been leaked?	Value
Age	\$0.0005
Gender	\$0.0005
ZIP Code	\$0.0005
Ethnicity	\$0.005
Education level	\$0.0005

Are you a millionaire?	Value
Yes	$0.116
No	$0

Are you engaged to be married? If so, how long?	Value
Yes, one month or less	$0.12
Yes, one to three months	$0.115
Yes, more than three months	$0.10
No	$0

Are you?	Value
Recently married?	$0.01
Recently divorced?	$0.01
Empty nester?	$0.01

What is your job?	Value
Accountant	$0.072
Altorney	$0.08
Banking and finance executive	$0.08
Chairman	$0.076
Chief executive	$0.086
Chief financial officer	$0.086
Chief information officer	$0.086
Chief operating officer	$0.086
Chief technology officer	$0.086
Company owner	$0.086
Cosmetologist+Beauty	$0.072
Entrepreneur	$0.10
Health professional	$0.072
Human resources executive	$0.08
Home improvement contractor	$0.072
Insurance agent	$0.072
Licensed professional	$0.072
Manufacturing & Engineering	$0.072
Non-profit	$0.072
Pilot	$0.072
Pharmaceutical industry exec	$0.076
President	$0.086
Real estate agent or broker	$0.072
Vice chairman	$0.086
Other	$0

A.2 Family & Health

Do you have children?	Value
Yes	$0.005
No	$0

Are you expecting a baby? If so, will this be your first childand which trimester are you in?	Value
Yes Yes First	$0.095
Yes Yes Second	$0.115
Yes Yes Third	$0.115
Yes No First	$0.08
Yes No Second	$0.10
Yes No Third	$0.10
No	$0

Are you a new parent? If so, is your new baby a boy or girl?	Value
Yes Boy	$0.035
Yes Girl	$0.035
No	$0

Do you have any of the following conditions?	Value
Acid reflux	$0.26
ADHD	$0.26
Allergies	$0.26
Arthritis	$0.26
Asthma	$0.26
Back pain	$0.26
Clinical depression	$0.26
Diabetes	$0.26
Frequent heartburn	$0.26
Headaches/migraines	$0.26

A.3 Property

Do you own a home?	Value
Yes	$0.085
No	$0

If you own a home, it is likely that data companies already know this information about you from public databases:	Value
The size of your home	$0.005
The size of your mortgage	$0.005
How many bathrooms the property has	$0.005
How many bedrooms the property has	$0.005

If you own a home, Has the following information been released on public databases?	Value
Yes	$0.085
No	$0

Is there a fireplace in your home?	Value
Yes	$0
No	$0

A.4 Activities

Do you have any of these hobbies?	Value
Are you a cruise enthusiast?	$0.03
Are you a fitness and excercise buff?	$0.03
Are you interested in foreign travel?	$0.03

Do you own an aircraft?	Value
Yes	$0.085
No	$0

Do you own a boat?	Value
Yes	$0.076
No	$0

Do you exercise or participate in other activities to lose weight?	Value
Yes	$0.105
No	$0

A.5 Consumer

Have you searched online or visited websites recently on any of these topics? (please select as many as appropriate)	Value
Auto	$0.0021
Financial information	$0.001
Retail	$0.001
Travel	$0.001
Gossip	$0.0013
Gaming	$0.0013
Food	$0.0013
Education	$0.0013
Cooking topics	$0.0008
Movie information	$0.003
Political and govermental topics	$0.0019
Telecom and television purchase research	$0.0015

Do you hold any store loyalty cards, at a grocery store or pharmacy, for instance?	Value
Yes	$0.001
No	$0

Are you looking to buy any of these products? (Select as many as appropriate)	Value
Car(s)	$0.0018
Consumer packaged goods such as soap, shampoo, toilet paper etc	$0.001
Education	$0.0013
Financial products or services	$0.001
Other vehicles	$0.0011
Clothes	$0.0008
Travel	$0.0011

Are you looking to buy a mobile phone?	Value
Yes	$0.0125
No	$0

B Privacy Act

In 1974, Congress was concerned with curbing the illegal surveillance and investigation of individuals by federal agencies that had been exposed during the Watergate scandal. It was also concerned with potential abuses presented by the government's increasing use of

computers to store and retrieve personal data by means of a universal identifier - such as an individual's social security number. The Act focuses on four basic policy objectives:

- To restrict disclosure of personally identifiable records maintained by agencies.
- To grant individuals increased rights of access to agency records maintained on themselves.
- To grant individuals the right to seek amendment of agency records maintained on themselves upon a showing that the records are not accurate, relevant, timely, or complete.
- To establish a code of "fair information practices" that requires agencies to comply with statutory norms for collection, maintenance, and dissemination of records.