cryptography

MDPI

Article

Data Sharing Privacy Metrics Model Based on Information Entropy and Group Privacy Preference

Yihong Guo [1,2], Jinxin Zuo [1,2,*], Ziyu Guo [1,2], Jiahao Qi [1,2] and Yueming Lu [1,2]

[1] School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China
* Correspondence: zuojx@bupt.edu.cn

Abstract: With the development of the mobile internet, service providers obtain data and resources through a large number of terminal user devices. They use private data for business empowerment, which improves the user experience while causing users' privacy disclosure. Current research ignores the impact of disclosing user non-sensitive attributes under a single scenario of data sharing and lacks consideration of users' privacy preferences. This paper constructs a data-sharing privacy metrics model based on information entropy and group privacy preferences. Use information theory to model the correlation of the privacy metrics problem, the improved entropy weight algorithm to measure the overall privacy of the data, and the analytic hierarchy process to correct user privacy preferences. Experiments show that this privacy metrics model can better quantify data privacy than conventional methods, provide a reliable evaluation mechanism for privacy security in data sharing and publishing scenarios, and help to enhance data privacy protection.

Keywords: privacy metrics; data security; information entropy; privacy preference

Citation: Guo, Y.; Zuo, J.; Guo, Z.; Qi, J.; Lu, Y. Data Sharing Privacy Metrics Model Based on Information Entropy and Group Privacy Preference. *Cryptography* **2023**, *7*, 11. https://doi.org/10.3390/cryptography7010011

Academic Editors: Hui Zhu and Yuan Zhang

Received: 19 November 2022
Revised: 1 March 2023
Accepted: 2 March 2023
Published: 3 March 2023

1. Introduction

With the development of the mobile internet, services such as big data and cloud computing, distributed computing, and storage services are gradually becoming more popular. Big data-enabled service delivery models such as user portraits [1] and swarm intelligence [2] were born. The service provider obtains the data and resources through the massive user terminal equipment and uses the client privacy data to carry on the machine learning and data mining work, improving service competitiveness and user experience. This leads to privacy data being out of the control of the original users in the service platform and devices for the collection, management, analysis, display, and so on [3], bringing new data security risks. The main research focuses on designing a good privacy protection algorithm, reducing the privacy of business processes, and scientifically quantifying data privacy.

The existing work on privacy metrics for data publishing relies on the artificial delineation of data scenarios and sensitivities to determine the modeling of privacy attacks. These methods only measure the leakage of sensitive attributes. For example, some works specify a patient's illness or an employee's income as sensitive attributes, thus calculating the leakage probability of sensitive attributes and the information loss of other identifiers or quasi-identifiers. In practice, users exposed to data tend to view all published personal information as sensitive and privacy-threatening, with only slight differences in the importance of attributes.

In order to deal with this situation, the current part of the work uses the entropy weight method to compute the privacy metrics. It gets the ranking of the privacy importance of different attributes of data. However, the traditional entropy weight method [4] is mainly used in the evaluation work and is not applicable in the privacy metrics scene. At the same

time, users in different scenarios and groups have different preferences for privacy because of individual conditions. Privacy metrics should analyze and incorporate differences in users' privacy preferences.

In response to the above problems, this paper makes the following contributions:

- In calculating the importance of privacy, we use information entropy to remodel the quantity of data privacy. After mathematical derivation, we use a new weight expression to replace part of the traditional entropy weight method.
- In the quantitative calculation of privacy preference, we add the analytic hierarchy process (AHP) [5] method to the data collection process and release and modify the results based on information entropy. The metric results fully take into account the user's personalized privacy preferences.
- We construct a complete data-sharing privacy metrics model, which provides a solution for evaluating privacy security in data-sharing scenarios. The experiments verify the validity of the model. Compared with the privacy metrics model based on the traditional entropy weight method, our model gets more reasonable weights and senses the change in data privacy more keenly.

The structure of the rest of this paper is as follows: The second part summarizes the current research situation on data privacy protection and privacy metrics. The third part introduces the framework of the privacy metrics model, the mathematical modeling of privacy metrics, and describes the privacy attribute weighting algorithm and the privacy preference modification process in detail. The fourth part describes the experimental process and results, and the fifth part summarizes and looks forward to the full text.

2. Related Work

According to the types and scenarios of data to be protected, privacy protection mainly includes location privacy, identifier anonymity protection, connection relationship anonymity protection, and so on [3]. Currently, the most commonly used privacy protection techniques are data anonymity based on generalization, data encryption based on cryptography, data disturbance based on noise, and their combination techniques [6]. K-Anonymity [7], l-Diversity [8], and t-Closeness [9] are the early privacy protection models based on the properties of the data itself. However, only anonymity is used to protect privacy, and the effect of de-anonymization is evident when the attacker has multiple information sources [10]. C. Dwork et al. [11,12] put forward a differential privacy protection model after strictly defining the background knowledge of the attacker.

Data is at the core of the internet of things, big data, and other services. Privacy protection in data publishing and sharing is an essential research issue in data security. Current data publishing privacy protection forms the architectural model [13] shown in Figure 1, which protects the privacy of data producers over the life cycle of data collection and sharing. M. H. Afifi et al. [14] presented a multi-variable privacy characterization and quantification model to provide metrics for data publishing and proposed distribution leakage and entropy leakage to better evaluate protection technologies. Abdelhameed S. A. et al. [15] proposed a restricted sensitive attributes based sequential anonymization (RSA-SA) approach. They introduced semantic and sensitivity diversity to measure and limit the privacy of published data. This method has a minor loss of information and delays time while preserving data privacy. J. Domingo-Ferrer et al. [16] redefined trust and data utility, tested them on a permutation model, and evaluated existing anonymization methods against new metrics, weighing information loss against the risk of privacy leakage. Z. G. Zhou et al. [17] proposed a re-anonymity architecture that released the generated Bayesian network rather than the data itself and optimized the excessive distortion of a specific feature attribute. Experimental results showed that this method could maintain privacy while maintaining high data availability.

Information theory is a vital information measurement tool that provides objective theoretical support in constructing the privacy metrics model and quantitative calculation. C. Díaz et al. [18] earlier applied information theory to privacy metrics, using information

theory to model privacy attacks in communication and changes in information entropy before and after attacks to measure the degree of anonymity of sensitive attributes. It is calculated and proven in several communication models. F. Gao et al. [19] used information theory to quantify privacy losses and trust gains in open, dynamic computing environments where private information is exchanged between trusted entities, the trust is dynamically adjusted to reduce the loss of privacy according to the situation of privacy leakage, and simulation experiments prove its effectiveness. Guizhou public big data key laboratory [20–22] has researched applying information theory to privacy metrics. Literature [20] puts forward a variety of information entropy universal models of privacy metrics from a theoretical point of view by assuming the attacker's existence and prior knowledge, as well as the subjective tendency of the user, in a well-conditioned privacy metric model that is gradually constructed. Literature [21] constructed a static game model with complete information between the service provider and the privacy attacker and modified the revenue matrix with the user's privacy preference. Finally, the mixed Nash equilibria with different preferences are obtained, and the privacy leakage in the process is measured using the strategy entropy. Literature [22] combined graph theory with information theory to construct a differential privacy metric model. The channel transfer matrix is transformed into a Hamming graph based on the graph's distance regularization and point transfer, and a metric method of privacy based on mutual information and differential privacy is proposed. It is proven that there is an upper bound for the amount of privacy leakage under differential privacy protection. Yu Yihan et al. [23] emphasized quantifying the privacy of the data itself and constructed an index of the elements of the privacy metric. On this basis, data privacy is quantified, and the entropy weight method is used to determine the weight of quantitative data. Finally, the BP neural network is used to complete the final classification of privacy. However, the process from data to a quantization matrix was relatively simple, and the relevance between the privacy metric and the information entropy model was not considered, which led to the distortion of the entropy weight method. Arcas S. et al. [24] questioned the application of information entropy in the measure of data anonymity, believing that data anonymity should be related to individuals and that the overall average amount of information in data tables measured by information entropy cannot fully reflect anonymity. Considering the uncertainty of the attacker's information and sensitive attributes, the method of information entropy is improved and compared with the conventional method on the data table. Zhao Mingfeng et al. [25] constructed a privacy metric model under swarm intelligence-aware scenarios, quantized the time series privacy data with non-negative mapping, and modified the user privacy preference matrix to obtain a privacy-sensitive data matrix by applying differential privacy protection to data. We observe the changing trend of data utility, privacy protection intensity, and privacy quantity, and the reliability of the metric model is proven.

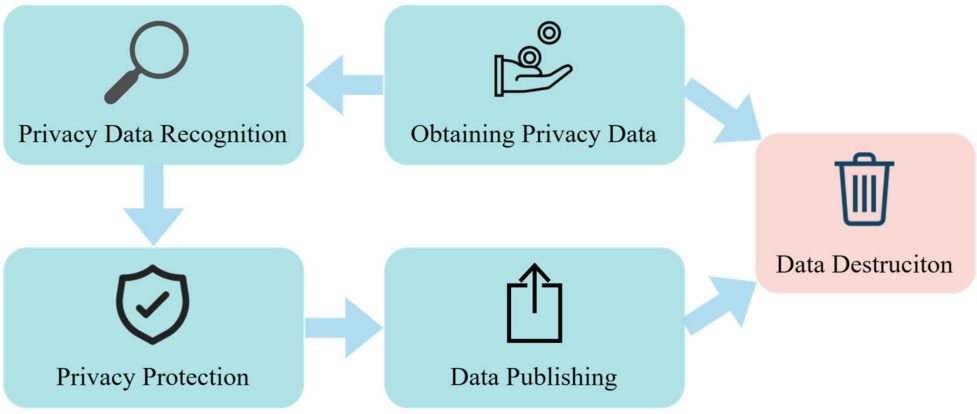

Figure 1. Privacy-preserving data publishing framework.

To sum up, the current research on privacy metrics has been fruitful, but there are still some problems to be solved: most of the scenarios modeled by some methods rely on theoretical assumptions, which are difficult to achieve in practice; The lack of an implementation framework for privacy metrics and protection as a whole; A large number of studies have focused on the privacy leakage of specific sensitive attributes in data. These methods are consistent with the actual situation in some scenarios, such as medical and finance. However, any individual information in the relationship will cause a certain degree of privacy leakage with only different attributes of privacy leakage, which are the importance of the degree of differences. The privacy metrics model needs to consider privacy preference, which is too subjective. Therefore, it is of great significance to construct a complete closed-loop model of privacy metrics and protection, to use good modeling and data processing methods to avoid distortion of calculation results, and to quantify users' privacy preferences from a group perspective.

3. Data Sharing Privacy Metric Model Based on Information Entropy and Group Privacy Preference

In order to solve the problems of poor correlation between the privacy metric method and actual data and the distortion of the calculation process, we construct a data-sharing privacy metric model based on information entropy and group privacy preference, as shown in Figure 2. The model divides the privacy metric process into three modules: problem modeling, quantitative calculation, and analysis-based decision-making. It guides privacy protection and data release decision-making, which ensures that privacy protection and metric linkage are closely combined.

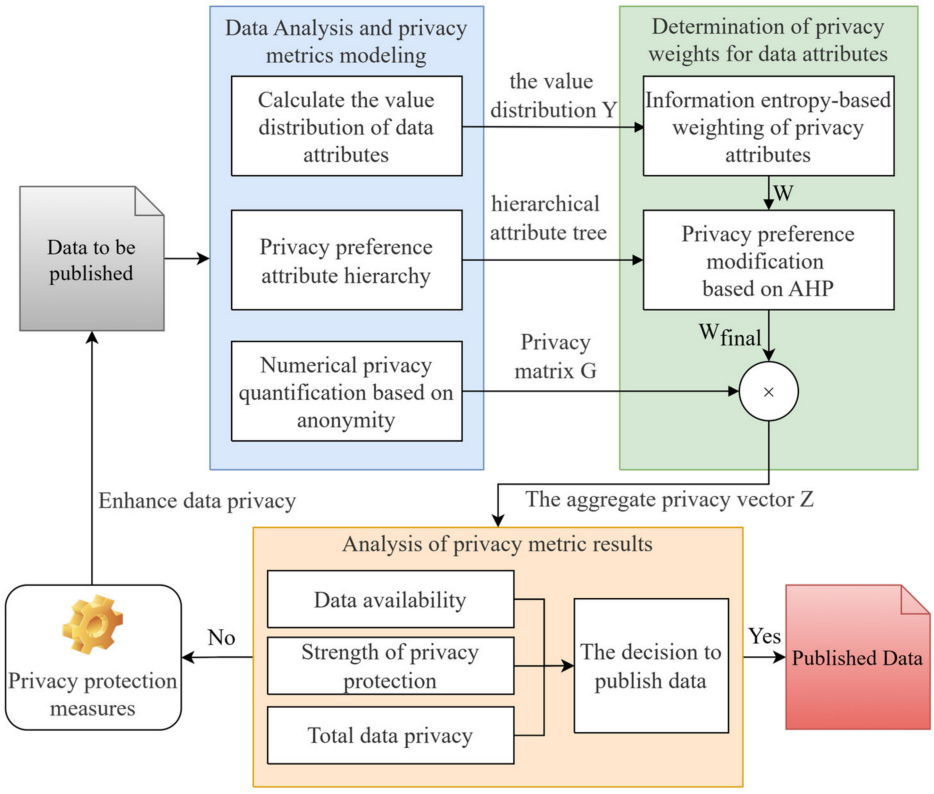

Figure 2. Data sharing privacy metrics model based on improved information entropy and group privacy preference.

First, when the privacy metric is applied in the data sharing and publishing scene, the model analyzes the statistical characteristics of the data based on the original data and obtains the distribution Y of different values. A hierarchical model of tree-type attributes is constructed based on the privacy preference statistics of user groups, and anonymity

is a primary feature to measure the strength of privacy protection, which is essentially a mathematical feature of the data itself. It is independent of the privacy protection algorithm and has some generality in its measurement. Then, the model quantizes the original data based on anonymity and obtains a computable fundamental privacy matrix G.

Additionally, the attribute privacy weights and the total privacy amount are calculated based on data analysis and privacy metric modeling results. The objective privacy attribute weight W is obtained by calculating the information entropy and its weight through the value distribution Y of data. Then, after calculating the group privacy preference vector p_{final} based on AHP, modify the subjective preference of W to obtain the final weight W_{final}. Using W_{final} and G to do weighted aggregation of the privacy metric, we get the integrated user privacy vector Z corresponding to the user one by one, that is, the user-level privacy metric result.

Finally, according to the application scenario, the total privacy metric results, the service data availability requirements, and the strength of the privacy protection algorithm are integrated to evaluate the data-sharing decision. Suppose that the decision-maker does not consider the data to meet the privacy requirements. In this case, the model will iteratively process the data, apply the corresponding privacy protection measures, and enhance the privacy protection of the data until the data meet the release conditions. The implementation details of the critical steps in the measurement process are detailed below.

3.1. Problem Modelling

Shannon information entropy is a measure of source uncertainty in the communication system. Suppose the state of system X is represented by a discrete random variable $X(x_1, x_2, \ldots, x_n)$, and $p(x_i), i \in [1,\ n]$ is the probability of each random event x_i occurring. For event x_i, by guesswork and proof of uniqueness, the amount of self-information that defines the occurrence of the event is

$$I(x_i) = \log\left(\frac{1}{p(x_i)}\right). \tag{1}$$

The lower the probability of the event, the higher the uncertainty of its elimination, and the greater the amount of self-information it contains. By calculating the mathematical expectation of the random variable X, the information entropy $H(X)$, which means the average uncertainty of the system, is obtained. The higher the uncertainty of the system, the greater the information entropy is

$$H(X) = \sum_{i=1}^{n} p(x_i) \log\left(\frac{1}{p(x_i)}\right) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)). \tag{2}$$

For data table T, there are n rows of private data associated with the user's identity. In previous data privacy metrics studies [13,23], models were often based on assumptions about attacks. Suppose the probability of an attacker recognizing a piece of information follows the distribution of the random variable X. In the initial state, since the attacker has no prior knowledge, every piece of information in T is assumed to be the target data with equal probability, that is, $x_i = \frac{1}{n}, i \in [1,\ n]$. The maximum of the uncertainty of X calculating by information entropy is H_0.

$$H_0 = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)) = -\sum_{i=1}^{n} \frac{1}{n} * \log\left(\frac{1}{n}\right) = \log(n) \tag{3}$$

Then the attacker obtains a value for one attribute a of the target to be identified as r_i and initiates a query Q on T. During the query, the probability of the attacker identifying each piece of data, that is, the probability distribution of X, changes and becomes X'. Suppose that the result of Q is n', that is, the number of data whose attribute has a value of r_i in T is n', and the new information entropy of X' is defined as

$$H_1 = -\sum_{j=1}^{n'} p(x_j|a = r_i)\log(p(x_j|a = r_i)) = -\sum_{i=1}^{n'} \frac{1}{n'} \times \log\left(\frac{1}{n'}\right) = \log(n'). \quad (4)$$

The change in information entropy before and after Query Q, $H_1 - H_0$, is taken as the measure value P of privacy leakage.

$$P = H_0 - H_1 = \log(n) - \log(n') = \log\left(\frac{n}{n'}\right) = \log\left(\frac{1}{n'/n}\right) \quad (5)$$

On this basis, let the random variable $Y = \{y_1, y_2, \ldots, y_k\}$ denote the probability distribution of attribute a getting a different value r_i in T, and there are k different values on attribute a. y_i represents the event whose attribute a values r_i, and p_i represents the probability that the event y_i occurs, which equals to n'/n. At this point, the mathematical representation of the self-information of the event y_i is consistent with P.

Suppose that the probability distribution of attribute a in T is the same as that of external data when the amount of data in T is large enough, and the attacker acquires prior knowledge from external data that contains the value of the target attribute a that can be associated with an attack on data table T. At this point, the average privacy leakage $\overline{P}$ of the correlation attack on T using attribute a is consistent with the mathematical expression of information entropy of random variable Y, as shown in Formula (6).

$$\overline{P} = H(Y) = \sum_{i=1}^{k} p(y_i)\log\left(\frac{1}{p(y_i)}\right) = -\sum_{i=1}^{k} p(y_i)\log(p(y_i)) \quad (6)$$

Based on the hypothesis and derivation above, it can be proved that the privacy weight of attribute a in data table T has the same mathematical expression as that of calculating information entropy from the probability distribution of value a; the latter can be used instead of the former. This conclusion provides a mathematical basis for the optimization of the following quantitative calculation process.

3.2. Weighted Algorithm for Privacy Attributes Based on Information Entropy

Based on the concept of information entropy in information theory, the traditional entropy weight method [5] is an algorithm to determine the importance of evaluation indeces by directly using the statistical characteristics of the target to be evaluated. The discreteness of the value distribution of numerical data determines the result of determining weights. After standardizing the data, the higher the degree of numerical discreteness under a single index, the smaller the result of the information entropy calculation is. Indicators that can distinguish data more effectively will obtain higher-weighted results. The privacy metric using the entropy weight method [23] can calculate the privacy quantity in the attribute as a weight. The calculation process of the traditional entropy weight method is as follows:

Firstly, Formulas (7) and (8) are used to standardize the statistical data of the evaluation target. Suppose that the evaluation data contains k indexes and n evaluation objects, where x_{ij} represents the value of the jth index in the ith target. Positive or negative is when the value of the index is positively correlated with the score of the evaluation result. The index is positive and standardized using Formula (7), standardization using Formula (8).

$$x'_{ij} = \frac{x_{ij} - \min\limits_{j}\{x_{ij}\}}{\max\limits_{j}\{x_{ij}\} - \min\limits_{j}\{x_{ij}\}} \quad (7)$$

$$x'_{ij} = \frac{\max\limits_{j}\{x_{ij}\} - x_{ij}}{\max\limits_{j}\{x_{ij}\} - \min\limits_{j}\{x_{ij}\}} \quad (8)$$

Using the standardized data, the corresponding entropy values for each index are obtained, as shown in Formula (9).

$$\begin{aligned} p_{ij} &= \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} (i = 1, 2, \ldots, n; j = 1, 2, \ldots, k) \\ H_j &= -\left(\frac{1}{\ln(n)}\right) \sum_{i=1}^{n} p_{ij} \log\left(p_{ij}\right) \end{aligned} \tag{9}$$

If $p_{ij} = 0$, define $\lim_{p_{ij} \to 0} p_{ij} ln p_{ij} = 0$. The calculated result is the normalized entropy value, $H_j \in [0, 1]$. Then the final index weight is obtained by using Formula (10), and the weight is inversely proportional to the calculated result of the entropy value.

$$w_j = \frac{1 - H_j}{k - \sum_{j=1}^{k} H_j} \ (j = 1, 2, \ldots, k) \tag{10}$$

The entropy weight method is an objective weight determination algorithm dependent on the data. It is suitable for the comprehensive evaluation process with many sampling targets. However, in data privacy metrics, there are the following problems when using the traditional entropy weight method directly after digitizing raw data:

- The entropy weight method uses the formula for information entropy, but the physical meaning of information entropy needs to be clarified. It lacks an explanation of scene modeling and probability angles, so it cannot be directly equivalent to privacy quantity.
- In the process of calculating privacy metrics, it is necessary to first quantify the privacy data and then calculate based on the quantized data. There are apparent differences between the data and the original data after privacy quantization and standardization, and improper data processing will lead to the distortion of entropy weight calculation.

Given the above problems, this paper improves the weight calculation process of the classical entropy weight method in privacy metrics. Section 3.1 is used to model the privacy attack on the original data and analyze its probability characteristics, based on which the mathematical derivation is carried out, using Formula (6) instead of the calculation process in the entropy weight method, as follows:

1. According to the original data, the random event $Y_j = \left(y_{1j}, \ldots, y_{mj}\right)$ of every index value is constructed, m is the number of the jth attribute that contains the value type, and the probability distribution $P\left(Y_j\right) = \left(p_{y_{1j}}, \ldots, p_{y_{mj}}\right)$ of every index value is calculated.
2. For each Y_j, the information entropy $H\left(Y_j\right) = -\sum_{i=1}^{m} p\left(y_{ij}\right) \log\left(p\left(y_{ij}\right)\right)$.
3. Since it has been proved in 3.1 that there is a directly proportional relationship between the amount of privacy leakage and the results of the current information entropy calculation, the final weight vector $w = \left(w_j\right), j = 1, 2, \ldots, k$ is obtained directly by $w_j = \frac{H_j}{\sum_{j=1}^{k} H_j}$ normalization.

Compared with the conventional numerical and entropy methods, the proposed method adds the modeling of the privacy metric scenario and makes the measurement process more interpretable. Based on the mathematical expression, the classical entropy weight method is adjusted and optimized to avoid distortion of the result and deviation from the expectation.

3.3. Weight Correction Based on User Privacy Preferences

In the existing data-sharing scenarios, there are some differences in data-sharing patterns, shared data types, user-oriented groups, etc. The demand for privacy protection is affected by subjective factors. The methods of determining users' privacy preferences in existing studies usually specify the rating directly [21], which does not consider the users' group wishes and may lead to the need for more integrity in the results of preference weighting. Privacy preferences are calculated for each user [25], the results are calculated

and retained, and there may be resource constraints on edge computing devices. AHP is a subjective, multi-criteria decision-making method. After constructing the hierarchy index system, the whole index importance ranking can be obtained by combining the subjective evaluation opinions of experts or users. AHP is applied to determine the individual privacy preference of group users. The steps of the algorithm are as follows:

1. Split the set of data attributes to be published and build the hierarchical attribute architecture. Building a hierarchical structure can not only help get more discriminative results when getting user privacy preferences but also avoid computing the eigenvalues of large matrices and improve the algorithm's efficiency. The privacy information category classifies the current attribute set and constructs the hierarchical model. For example, 17 personal data attributes can be split and built into an attribute hierarchy, as shown in Figure 3.

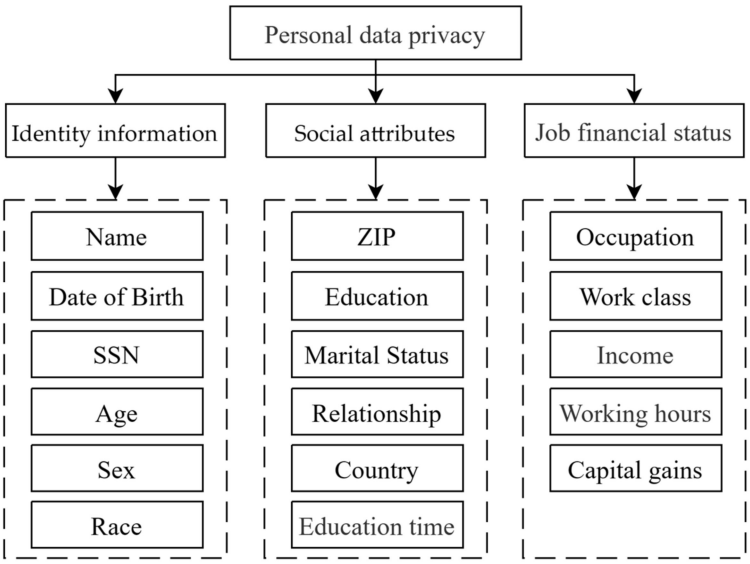

Figure 3. Hierarchy Construction of Personal Privacy Information Attributes.

The current hierarchy of data attributes is based on the following rules:

- Identity information: attributes related only to the individual, independent of others, natural attributes of the person.
- Social attributes: the attributes that describe individual participation in social relations; attributes related to others; and regional attributes.
- Job financial status: an attribute that describes an individual's occupation and financial status.

2. Based on the hierarchical model, the relative importance of users' privacy preferences is analyzed, and the judgment matrix $C_{t\times t} = [b_{ij}]_{t\times t}$ is constructed for each sub-level, assuming that the number of attributes in the sub-level is t. Table 1 is a quantitative representation of subjective opinions, which quantifies abstract and fuzzy user opinions into a numerical matrix by comparing two different attributes. For each b_{ij} in the matrix, using the scale in Table 1, get the numerical expression of the user group's preference for attributes by pairwise comparison.
3. Single-level sorting. The method of square root or sum product is used to calculate the maximum eigenvalue $\lambda_{\max}$ of matrix $C_{t\times t}$ and its corresponding eigenvector p.
4. Consistency test. There may be some conflicts between two sets of comparisons, and consistency needs to be verified to ensure the validity of the statistics. Since the data is transformed to the judgment matrix $C_{t\times t}$, the problem is transformed to determine whether the matrix $C_{t\times t}$ is consistent, that is, whether the largest eigenvalue $\lambda_{\max}$ of the matrix equals the order of the matrix t. However, absolute consistency is often challenging to achieve, so the use of an approximate way to measure the degree

of consistency of the matrix at this time. To avoid the inconsistency caused by the statistics of subjective privacy preferences, the consistency test should be carried out on the calculated results. The consistency index $C.I.$ was obtained by using Formula (11). The random consistency index $R.I.$ is selected according to Table 2 and index number t.

Table 1. Scale Method.

Intensity of Importance	Definition	Explanation
1	Equal importance	The degree of contribution of the two elements is of equal importance
3	Moderate importance of one over another	Experience and judgment slightly prefer the former
5	Essential importance	Experience and judgment strongly prefer the former element
7	Extreme importance	Actually shows a very strong preference for the former element
9	Absolute importance	There is sufficient evidence to confirm an absolute preference for the former element
2, 4, 6, 8	Intermediate value of adjacent scale	Between two adjacent judgments
Reciprocals	Relative unimportance	The degree of the latter factor preference is inversely proportional to the value, and the smaller the value, the higher the importance of the latter.

Table 2. Random consistency index.

t	*R.I.*	T	*R.I.*
1	0	9	1.46
2	0	10	1.49
3	0.52	11	1.52
4	0.89	12	1.54
5	1.12	13	1.56
6	1.26	14	1.58
7	1.36	15	1.59
8	1.41		

$$C.I. = \frac{\lambda_{\max} - t}{t - 1} \tag{11}$$

$C.I.$ is of the following nature:

- The matrix C has complete consistency when $C.I. = 0$.
- When $C.I.$ is close to zero, the matrix C has satisfactory consistency.
- The greater the $C.I.$, the greater the inconsistency of C.

Finally, calculating the conformance ratio $C.R. = C.I./R.I.$ In general, the result passes the conformance test when $C.R. < 0.1$. The judgment matrix obtained from each user's relative importance survey must satisfy the consistency test condition.

After passing the consistency test, the weight vector p is the user's privacy preference weight for the index. After a survey of multiple users, all the results are weighted equally to get the final target group's privacy preference weight, p_{final}.

By determining a suitable proportion coefficient α and β, the information entropy weight vector w_{final} with the group privacy preference is obtained by modifying w with p_{final}, as shown in Formula (12).

$$\begin{gathered} w_{final} = \alpha w + \beta p_{final}, \alpha + \beta = 1 \\ \alpha - \beta = Co\left(w, p_{final}\right) = \sqrt{\frac{\sum_{j=1}^{m}\left(w - p_{final}\right)^2}{2}} \end{gathered} \tag{12}$$

where $Co(w_a, w_b)$ represents the correction function. The value of $Co(w_a, w_b)$ satisfies the normalization condition. The principle of weight allocation is that the information entropy-based privacy metrics (objective weights) be appropriately modified by the user's privacy preferences (subjective weights) without affecting the dominant position of the objective weights. The distance of the weight and the distance of the weight coefficient are consistent, and the weight distribution is more objective and reasonable. When the difference between the two sides is large, the subjective information introduced by the user's privacy preference is limited. When the difference between the two sides is small, the correction effect of the user's privacy preference is reflected as much as possible.

In order to better adapt the AHP method in the work of privacy measurement and protection, our scheme tries to add AHP into the process of privacy measurement of data publication, forming a process of data publishing privacy metrics as shown in Figure 4, which integrates the AHP method. The process consists mainly of the following steps:

- By describing a problem that needs data support, the data consumer puts forward the demand for data usage and sends the demand to the data server.
- According to the requirement, the data server formulates the data attributes that need to be collected, divides the data attributes according to specific rules, and constructs the hierarchical structure model.
- The data server requests that the user group use the actual data. The user gives the privacy importance preference matrix about the data attribute and sends the preference opinion to the data server.
- The data server integrates the data and preferences of each user individually to obtain the actual original data to be published and the group privacy preference matrix.
- The data server iterates through the privacy metrics and protection model shown in Figure 3 to get the data that meets the privacy requirements.
- The data server publishes the final data and provides it to the data consumer for analysis and sharing.

According to the above process, the data server can better integrate the AHP method into the data collection and release process and develop the application of user group privacy preferences in privacy measurement and protection.

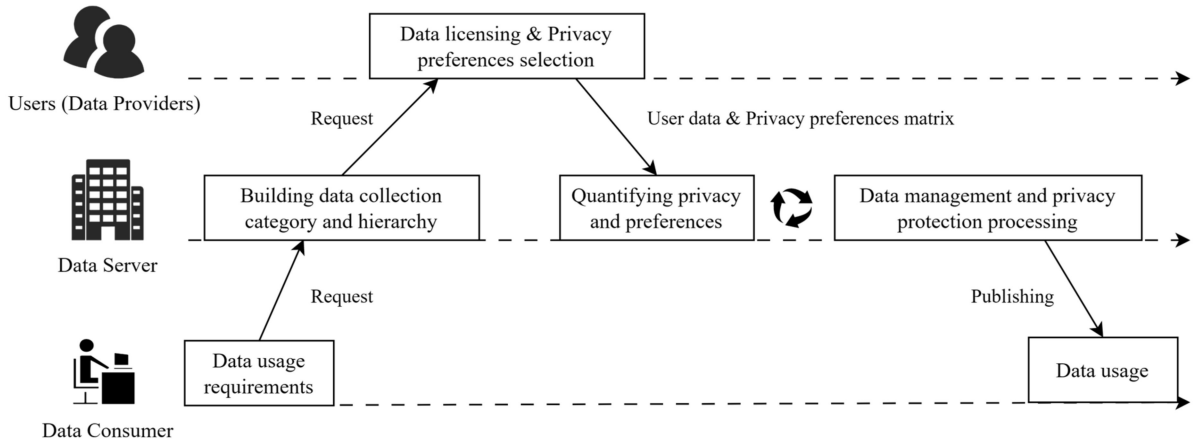

Figure 4. Data release and privacy metric process after joining AHP.

3.4. Metric Results Analysis and Feedback

The feedback from measurement results analysis is an essential part of the whole work of privacy metrics and privacy protection. In this link, according to the calculation results of privacy metrics and the business environment, the model makes the data release and privacy adjustment decisions. When sharing or publishing data, consider the following factors:

- External environmental factors. Business scenarios for data usage and the network environment for data transmission. It will dynamically influence the security requirements for data sharing and circulation and restrict the adoption and strength of data security and privacy protection.
- Data source privacy. The privacy attributes, information, and statistical characteristics of the original data source are mainly determined by the privacy metrics mentioned above.
- Data availability. Data that has been protected after processing should be guaranteed to be available. Consider the destruction of crucial information in the data, the destruction of the original distribution, and so on. Protection of privacy and security at the same time, as far as possible, to minimize the impact of protection measures on data utility.

We can judge whether the current data can be released or not by the above factors. Suppose it does not meet the privacy and data availability requirements in the current situation. In this case, we need to adjust the privacy protection measures dynamically, such as by changing the applicable privacy protection algorithm or adjusting the strength of the privacy protection algorithm, and iteratively measure the processed data again until the data meet the release requirements after analysis and implementation of the data release decision.

4. Experimental Results and Discussion

This paper simulates non-interactive data publishing and user preference groups with social network scenarios and unpublished data sources using adult data sets provided by UCI [26]. The data set contains 19 attributes of the user's personal information. We filtered and merged the attribute sets to get 17 attributes and selected the first 1000 user personal information as the privacy metrics of the data to be published.

4.1. Comparative of Weight Distribution of Data Privacy Attributes

The reasonable weight of privacy attributes is the basis for obtaining the privacy of each data item and is the guarantee of the scientific results of the whole privacy quantification calculation. For the raw data without privacy protection, different exact weight algorithms are used to get the weight vector, and the weight distribution of each method is compared, as shown in Figure 5.

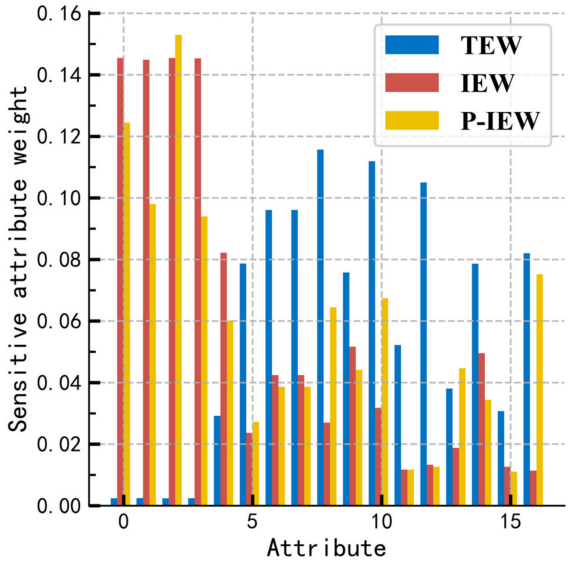

Figure 5. Comparative analysis of attribute privacy weights of different methods.

The blue column represents the conventional data frequency quantization with entropy weight (TEW) [23], and the red column represents the information entropy quantization

weight (IEW) without privacy preference modification in the scheme. The yellow bars represent information entropy quantization weights (P-IEW) with privacy preference modifications. The original data set attributes 1~4 are direct identifiers, or the high distinguishing ability of the standard identifier attributes, such as name, SSN, etc., should have a high privacy weight. The original method uses the entropy weight method for the data after the numerical frequency quantization and lacks the scientific modeling explanation, and the quantized data no longer has the statistical characteristics of the original data, which leads to a large deviation between the weight and the expectation. This scheme strictly obeys the modeling of the privacy metric problem, optimizes the weighting algorithm under the premise of satisfying the mathematical interpretation of information entropy, and accords with the expectation in the data of the privacy attribute weight calculation, so the problem of weight distortion of the original method is solved.

Concurrently, the statistical data of users' subjective privacy preferences in social network scenarios is obtained based on simulation experiments. Each constructed judgment matrix passed the consistency verification condition. After calculating by the AHP method, the privacy preference of information entropy weight is modified by weighted aggregation. The adjusted weight has a higher sensitivity, a higher attribute discrimination degree, and a lower objective weight, such as family relationship and personal income. Still, the user group's subjective weight of privacy information is compensated. The results of the privacy metrics fit the expectations of the users of the simulated social network group privacy preferences and make the results more suitable for the needs of the scenario.

4.2. Measures of Privacy Protection Effectiveness

A reasonable measure of data privacy needs to give relatively straightforward quantitative feedback after handling the data with the protective measures, which reflects the effect of the protective measures and then guides the control and adjustment of privacy protection. By adding privacy protection measures to the original data, the experiment further verified the validity and superiority of this scheme under the privacy protection treatment. Figure 6 shows the trend of the model's data privacy metrics results after suppressing the identifier attributes individually. By analyzing the maximum value of individual privacy and the average value of total privacy, we can find that with the increase in suppression of attributes, the effect of privacy protection is enhanced. The amount of data privacy continues to decrease, which is inversely proportional to the strength of privacy protection. The trend of change is following the expectation of experience, which proves the validity of the measurement effect.

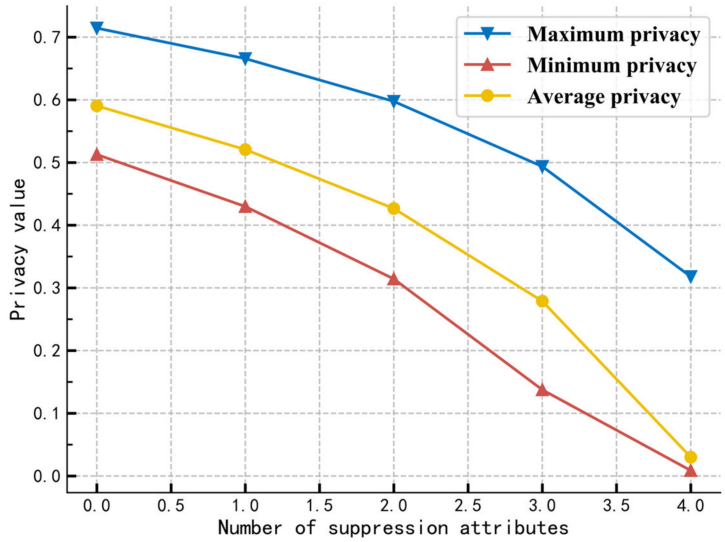

Figure 6. Suppression-overall privacy metrics trends.

Further, performing a visual analysis of the changes in the distribution of individual data privacy during attribute suppression, as shown in Figure 7. Individual privacy fluctuates around the average of group privacy. With the increase in privacy protection intensity, the mean of individual privacy decreases and the fluctuation amplitude of individual privacy increases. The metric results help provide privacy decision-makers with a basis for classified protection on two dimensions:

- Classification of attributes: providing classification protection according to the sensitivity of attributes and the identification ability of individuals. It can guarantee data availability on low-sensitive attributes and provide key protection for data on high-sensitive attributes by slicing, suppression, generalization, and so on.
- Classification of individuals: dividing all individuals in the relationship data into high, medium, and low areas according to the average privacy amount and individual sensitivity. Limit the release of highly sensitive individual data through permutation, bucket splitting, and perturbation techniques while reducing the overall privacy impact and providing high availability of data.

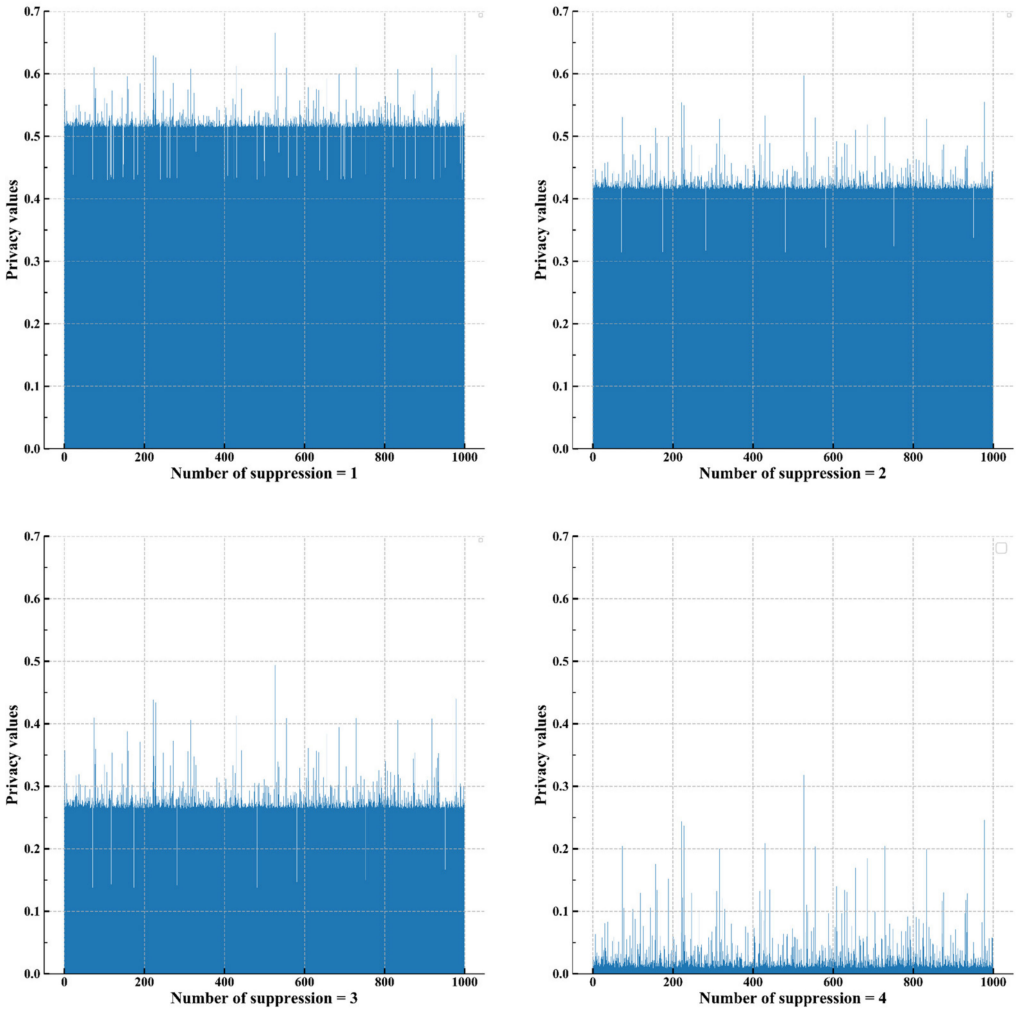

Figure 7. Suppression-individual privacy metrics distribution.

Figure 8 shows the effect of the three approaches on privacy measures when increasing the intensity of generalization applied to the data. The pretreatment eliminates some non-numerical attributes and obtains the numerical data by string processing and type conversion. To keep the same generalization strength ϵ of all the data, different generalization cardinality L is used for each attribute, and the generalization step S is used to control ϵ, and process the generalized data.

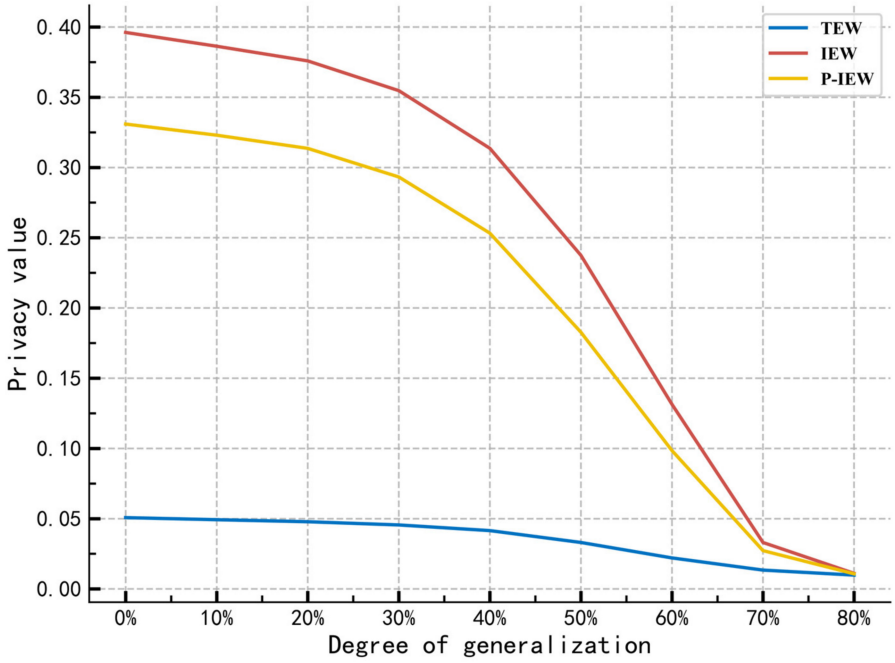

Figure 8. Generalization—metric sensitivity analysis.

Through the comparison, we can see that the privacy attribute weights of IEW and P-IEW calculated by the method of determining the privacy attribute weights in the scheme of this paper, and the total privacy of the final calculation, be able to measure changes in data privacy more sensitively.

$$K = \frac{1}{n-1} \sum_{i=1}^{n-1} \left| \frac{y_i - y_{i+1}}{x_i - x_{i+1}} \right| \tag{13}$$

More specifically, using the average change rate K of the privacy measures shown in Formula (13) as an index to evaluate the sensitivity of the above privacy measure scheme, the higher the K value, the higher the sensitivity of the corresponding scheme's privacy measure. Among them, $K_{TEW} = 0.051$, $K_{IEW} = 0.481$, $K_{P-IEW} = 0.400$, the latter two being higher than the traditional method. The results mean that the scheme can be more sensitive to changes in privacy intensity and data privacy quantity in the work of privacy protection and adapt to more detailed quantitative privacy analysis and intensity control.

Furthermore, the entropy attribute weights modified by user group privacy preferences are less sensitive than those before the modification. Due to the difference between the subjective privacy preference weight distribution and the probability distribution based on the objective information theory, users may protect some sensitive attributes that cannot easily cause privacy attacks. To reflect the subjective will of users, the model may sacrifice some privacy metrics in the quantization results. Therefore, in practice, it is necessary to dynamically select and adjust the influence of the user's subjective preferences on the measurement results to satisfy the requirement of measurement results.

5. Conclusions

This paper designs a privacy metrics algorithm based on information entropy and user privacy preferences to solve the problem of privacy metrics in the data-sharing publishing scene. This method considers all the data attributes to be published in the ranking of privacy importance and modifies the weights using the privacy preferences quantified by AHP. In addition, this paper proposes a privacy protection and metrics model for data sharing and publishing that simulates data sharing scenarios and various privacy protection measures using personal identity privacy data and verifies the generality and validity of the model. The model is suitable for quantitative analysis of data privacy in

the early days of data sharing. It can guide the privacy protection staff to implement the best data privacy protection measures according to the privacy protection threshold, the service data availability needs, the privacy classification, and the classification scheme. In the provision of data sharing, publishing, computing, and other services, to protect users' privacy as much as possible.

Author Contributions: Conceptualization, Y.G., J.Z. and Z.G.; methodology, Y.G. and J.Q.; validation, Y.G.; writing—original draft preparation, Y.G.; writing—review and editing, J.Z. and Z.G.; supervision, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Key R&D Program of China (No. 2021YFB3101903), the National Key R&D Program of China (No. 2022YFB3104900), and the National Key R&D Program of China (No. 2019YFB2102403).

Data Availability Statement: We used the Adult Data Set [24] for our experiments.

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ouaftouh, S.; Zellou, A.; Idri, A. User Profile Model: A User Dimension Based Classification. In Proceedings of the 10th International Conference on Intelligent Systems: Theories and Applications (SITA), Taipei, China, 17 December 2015.
2. Ganti, R.K.; Ye, F.; Lei, H. Mobile crowdsensing: Current state and future challenges. *IEEE Commun. Mag.* **2011**, *49*, 32–39. [CrossRef]
3. Feng, D.G.; Zhang, M.; Li, H. Big data security and privacy protection. *Chin. J. Comput.* **2014**, *37*, 246–258.
4. Saaty, T.L. Decision making with the analytic hierarchy process. *Int. J. Serv. Sci.* **2008**, *1*, 83–98. [CrossRef]
5. Zou, Z.H.; Yun, Y.; Sun, J.N. Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. *J. Environ. Sci.* **2006**, *18*, 1020–1023. [CrossRef] [PubMed]
6. Zhou, S.G.; Li, F.; Tao, Y.F.; Xiao, X.K. Privacy preservation in database applications: A survey. *Chin. J. Comput.* **2009**, *32*, 847–861. [CrossRef]
7. Sweeney, L. K-Anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
8. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-Diversity: Privacy Beyond K-Anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE), Atlanta, GA, USA, 3 April 2006.
9. Li, N.; Li, T.; Venkatasubramanian, S. T-Closeness: Privacy beyond k-Anonymity and l-Diversity. In Proceedings of the 23rd International Conference on Data Engineering (ICDE), Istanbul, Turkey, 14 May 2007.
10. Zang, H.; Bolot, J. Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study. In Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, MOBICOM 2011, Las Vegas, NV, USA, 9 September 2011.
11. Dwork, C. Calibrating noise to sensitivity in private data analysis. *Lect. Notes Comput. Sci.* **2006**, *3876*, 265–284.
12. Dwork, C.; Roth, A. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* **2013**, *9*, 211–407. [CrossRef]
13. Qu, L.; Yang, J.; Yan, X.; Ma, L.; Yang, Q.; Han, Y. Research on Privacy Protection Technology for Data Publishing. In Proceedings of the 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), Hainan, China, 6–10 December 2021; pp. 999–1005.
14. Afifi, M.H.; Zhou, K.; Ren, J. Privacy Characterization and Quantification in Data Publishing. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1756–1769. [CrossRef]
15. Abdelhameed, S.A.; Moussa, S.M.; Khalifa, M.E. Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing. *Knowl.-Based Syst.* **2019**, *164*, 1–20. [CrossRef]
16. Domingo-Ferrer, J.; Muralidhar, K.; Bras-Amoros, M. General Confidentiality and Utility Metrics for Privacy-Preserving Data Publishing Based on the Permutation Model. *IEEE Trans. Dependable Secur. Comput.* **2021**, *18*, 2506–2517. [CrossRef]
17. Zhou, Z.; Wang, Y.; Yu, X.; Miao, J. A Targeted Privacy-Preserving Data Publishing Method Based on Bayesian Network. *IEEE Access* **2022**, *10*, 89555–89567. [CrossRef]
18. Dıaz, C.; Seys, S.; Claessens, J.; Preneel, B. Towards Measuring Anonymity. In Proceedings of the 2nd International Workshop on Privacy-Enhancing Technologies, San Francisco, CA, USA, 14 April 2002.
19. Gao, F.; He, J.; Peng, S.; Wu, X. A Quantifying Metric for Privacy Protection Based on Information Theory. In Proceedings of the 3rd International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, China, 4 February 2010.

20. Peng, C.G.; Ding, H.F.; Zhu, Y.J.; Fu, Z.F. Information entropy models and privacy metrics methods for privacy protection. *J. Softw.* **2016**, *27*, 1891–1903.
21. Zhang, P.P.; Peng, C.G.; Hao, C.Y. Privacy protection model and privacy metric methods based on privacy preference. *Comput. Sci.* **2018**, *45*, 130–134.
22. Wang, M.N.; Peng, C.G.; He, W.Z.; Ding, X.; Ding, H.F. Privacy metric model of differential privacy via graph theory and mutual information. *Comput. Sci.* **2020**, *47*, 270–277.
23. Yu, Y.H.; Fu, Y.; Wu, X.P. Metric and classification model for privacy data based on Shannon information entropy and BP neural network. *J. Commun.* **2018**, *39*, 10–17.
24. Arca, S.; Hewett, R. Is Entropy Enough for Measuring Privacy? In Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA,, 16 December 2020.
25. Zhao, M.F.; Lei, C.; Zhong, Y.; Xiong, J.B. Dynamic privacy measurement model and evaluation system for mobile edge crowdsensing. *Chin. J. Netw. Inf. Secur.* **2021**, *7*, 157–166.
26. Kohavi, R.; Becker, B.; University of California. Adult Data Set. UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ml/datasets/Adult (accessed on 25 July 2022).