Privacy and Data Utility Measurement Model for Structured Data (Postprint)

Authors: Xie Mingming, Peng Changgen, Wu Ruixue, Ding Hongfa, Liu Botao

Date: 2019-04-01T00:00:00+00:00

Abstract

To address the quantification problem of data privacy amount and data utility in privacy protection, we propose a structured data privacy and data utility measurement model based on the fundamental principles of metric spaces and norms. First, we present a data numerical processing method that transforms data tables into matrices for computation; second, we introduce a privacy preference function to measure the variation of sensitive attributes over time; then, we analyze the privacy protection model to quantify the changes produced by privacy protection techniques; finally, we construct a metric space and provide calculation formulas for privacy amount, data utility, and privacy protection level. Through example analysis, the proposed measurement model can effectively reflect the amount of privacy information.

Full Text

Preamble

Privacy and Data Utility Metric Model for Structured Data

Xie Mingminga,b, Peng Changgenb,c†, Wu Ruixuec,d, Ding Hongfaa,b, Liu Botaob,c

(a. College of Mathematics & Statistics; b. Guizhou Province Key Laboratory of Public Big Data; c. College of Computer Science & Technology; d. Institute of Cryptography & Data Security, Guizhou University, Guiyang 550025, China)

Abstract: Addressing the quantification challenges of data privacy and data utility in privacy protection, this paper proposes a privacy and data utility metric model for structured data based on the fundamental principles of metric spaces and norms. First, a data numerical processing method is presented to transform data tables into matrices for computation. Second, a privacy

preference function is introduced to measure the temporal variation of sensitive attributes. Then, privacy protection models are analyzed to quantify the changes produced by privacy protection techniques. Finally, a metric space is constructed, and formulas for calculating privacy amount, data utility, and privacy protection degree are derived. Instance analysis demonstrates that the proposed metric model can effectively reflect private information content.

Keywords: privacy protection; privacy metric; metric space; privacy amount; data utility

0 Introduction

We have entered the era of big data, where data permeates every industry and business function, becoming a critical production factor. In practice, many organizations must regularly release data—such as medical, transportation, and government data—which contains substantial personal privacy information. Disclosure of such data could cause immeasurable harm. In data publishing, to prevent complete public exposure of privacy data, organizations typically employ privacy protection techniques to conceal users' sensitive attributes. However, critical questions remain: Can processed data still leak privacy? How much privacy content exists? What impact does processing have on data utility? These factors are key to data publication. Without effective metrics for privacy and data utility, organizations face a dilemma: possessing data but fearing to release it, resulting in low degrees of open sharing and preventing effective value extraction. Thus, research on privacy metrics is urgently needed.

Privacy measurement methods fall into three categories: (1) probabilistic statistical methods that infer privacy leakage risk using probability distribution information; (2) information entropy methods that measure privacy information through uncertainty in information systems; and (3) set pair analysis theory, a qualitative-quantitative approach for addressing certainty-uncertainty problems.

Regarding probabilistic methods, Li et al. [1-2] proposed in 2007 and 2010 a metric based on k-anonymity and l-diversity that calculates sensitive attribute value distributions, using Earth Mover' s Distance (EMD) to compute divergence between global distribution of sensitive attributes in data and distribution within equivalence classes. Smaller divergence indicates lower privacy leakage risk. Since EMD doesn' t consider stability between equivalence classes and data distribution, Zhang et al. [3] proposed the EKM metric in 2014, combining EMD and KL divergence to measure privacy leakage risk through both distribution divergence and stability divergence. Based on probability distributions of sensitive attributes, references [4-6] between 2015-2017 proposed Bayesian inference-based privacy leakage metrics, analyzing differences between inferred and actual privacy information—smaller differences indicate higher leakage risk.

Information entropy, foundational to communication theory, quantifies information uncertainty. For privacy metrics using entropy, Díaz et al. [8] first applied information entropy to privacy protection in 2002, proposing its use for measuring anonymity in anonymous communication systems. In 2006, Clauß et al. [9] used entropy to describe uncertainty of privacy information in datasets. In 2007, Hoh et al. [10] measured trajectory tracking uncertainty based on entropy, proposing a novel temporal confusion metric for location trajectory privacy. In 2009, Ma et al. [11,12] adopted information-theoretic methods, quantifying location privacy as uncertainty linking location information to specific individuals. Shokri et al. [13] proposed a distortion-based privacy metric, reflecting user privacy levels by comparing differences between attackers' observed trajectories and users' actual trajectories. In 2011, Chen et al. [14] used conditional entropy to measure query privacy in Location-Based Services (LBS). In 2012, Yang et al. [15] identified individuals from sensitive information in network access, proposing two attacker types and using entropy to measure threats to general users. In 2016, Peng et al. [16] modeled privacy protection systems as communication models to make entropy-based metrics more intuitive, proposing several privacy protection information entropy models and theoretically deriving universal privacy metrics.

Set pair analysis theory [17] characterizes the deterministic-uncertain relationship of shared attributes between two connected sets by establishing connection numbers of identity, difference, and opposition. In 2015, Yan et al. [18] proposed a novel set pair analysis method for user privacy protection metrics, establishing privacy metric system standards and content for three applications: database privacy protection, location privacy protection, and trajectory privacy protection.

References [19-23,28-30] also describe and research privacy metrics, demonstrating that privacy information measurement methods continue evolving with increasingly comprehensive theoretical foundations, though more thorough investigation remains needed. To express privacy information content more intuitively from a mathematical perspective, this paper employs fundamental metric space theory to propose a privacy and data utility metric model for structured data.

---

1 Preliminaries

1.1 Metric Space

In mathematics, a metric space is a set where distance between any two elements is definable.

Definition 1 (Metric Space). Let R be a nonempty set whose elements are called points. For any two points $x, y \in R$, assign a real number $\rho(x, y)$ satisfying:

a) $\rho(x, y) \geq 0$, with $\rho(x, y) = 0$ iff $x = y$ (non-negativity and identity of indiscernibles)

b) $\rho(x, y) = \rho(y, x)$ for any points $x, y \in R$ (symmetry)

c) For any $x, y, z \in R$, $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (triangle inequality)

$\rho(x, y)$ is called the distance between x and y, and R equipped with distance ρ is called a metric space, denoted (R, ρ).

From this definition follow these properties:

d) Symmetry: $\rho(x, y) = \rho(y, x)$

e) For any $x, y, z \in R$, $|\rho(x, z) - \rho(y, z)| \leq \rho(x, y)$

1.2 Vector and Matrix Norms

Norms are fundamental concepts in functional analysis, commonly used to measure length or magnitude of elements in a space. Below we introduce norms for vector and matrix spaces.

For vector $x = (x_1, x_2, \ldots, x_n)^T$ and matrix $A = (a_{ij})_{m \times n}$:

- Vector 1-norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- Vector 2-norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$
- Matrix F-norm: $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$

1.3 Differential Privacy

Differential privacy is a data distortion-based privacy protection technique that adds random noise to sensitive data to achieve privacy protection while maintaining certain data utility. The principle is to add random noise making two datasets differing by at most one record indistinguishable, preventing individual privacy inference from query results.

Definition 2 (Differential Privacy). Let datasets D and D' have identical attribute structure, differing by at most one record. Let M be a randomized algorithm with range $Range(M)$. If for any output $O \subseteq Range(M)$, algorithm M satisfies:

$$\Pr[M(D) \in O] \leq e^{\varepsilon} \cdot \Pr[M(D') \in O]$$

then M satisfies ε-differential privacy, where ε is the privacy budget. Privacy protection degree is controlled by limiting ε: smaller ε means larger noise, higher privacy protection, but lower data utility; larger ε means smaller noise, lower privacy protection, but higher data utility.

For numeric data, Laplace distribution noise can provide ε-differential privacy. Let random variable x have probability density function:

$$f(x|\mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

where μ is location parameter, $b>0$ is scale parameter. Then x follows Laplace distribution $Laplace(\mu,b)$, denoted $x\sim Laplace(\mu,b)$. Its cumulative distribution function is:

$$F(x)=\begin{cases}\frac{1}{2}\exp\left(\frac{x-\mu}{b}\right), & x<\mu\\ 1-\frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right), & x\geq\mu\end{cases}$$

Its inverse cumulative distribution function is:

$$F^{-1}(p)=\mu-b\cdot\mathrm{sgn}(p-0.5)\cdot\ln(1-2|p-0.5|)$$

Laplace-distributed random numbers are generated via uniformly distributed random numbers and the inverse CDF to add noise satisfying differential privacy.

1.4 Variable, Symbol, and Terminology Definitions

1) Terminology:

a) **Data Utility** refers to the similarity or authenticity between processed data and original data from the same group. Higher authenticity means higher utility.

b) **Privacy Preference Timeliness** means an individual' s concern for a sensitive attribute changes over time. For example, a patient with "tumor" highly values disease privacy during illness but after recovery, disclosure of past "tumor" history becomes less concerning. Thus, concern for this sensitive attribute decreases over time.

2) Variables and Symbols:

For matrix operations, let A be an $m\times n$ matrix $A=(a_{ij})_{m\times n}$, abbreviated as $A=(a_{ij})$. Let B be a matrix with identical dimensions $B=(b_{ij})_{m\times n}$. Operations are defined as:

- Matrix addition: $A+B=(a_{ij}+b_{ij})_{m\times n}$
- Hadamard product: $A\circ B=(a_{ij}\cdot b_{ij})_{m\times n}$
- Scalar multiplication: $kA=(k\cdot a_{ij})_{m\times n}$

2 Privacy and Data Utility Metric Model

To measure privacy amount in structured data publishing and data utility after privacy protection processing, this paper constructs a metric model based on functional analysis principles. First, structured data is numerically processed to obtain a sensitive data matrix. Second, considering subjective privacy preferences and their timeliness, three types of privacy preference functions describe sensitivity variation. Third, privacy protection model impacts on the sensitive data matrix are analyzed and quantified. Finally, distances between sensitive data matrices are constructed to measure privacy and utility.

2.1 Data Numerical Processing

In structured data, each individual' s record attributes fall into four categories: explicit identifier attributes, quasi-identifier attributes, sensitive attributes (SA), and non-sensitive attributes (NA). Since explicit identifiers are directly removed and quasi-identifiers have standard numerical methods, while non-sensitive attributes fall outside privacy protection scope, this paper only processes sensitive attributes numerically.

Structured data with sensitive attributes is described as in Table 1 , where u_i represents the i-th individual, SA_j represents the j-th sensitive attribute, and $data_{ij}$ represents the j-th sensitive attribute value of the i-th individual.

Definition 3 (Non-negative Numerical Mapping). Let X be a finite set of non-numeric elements, $f : X \to \mathbb{R}$ a mapping. If for each $x \in X$, $f(x) \geq 0$, then f is a non-negative numerical mapping. The set of all such mappings is denoted $\mathcal{F}$.

Based on each sensitive attribute' s inherent sensitivity characteristics, following the principle that higher sensitivity maps to larger values, select n non-negative numerical mappings $f_1, f_2, \ldots, f_n \in \mathcal{F}$ to numerically process Table 1:

$$d_{ij} = f_j(data_{ij}), \quad i = 1, 2, \ldots, m; \ j = 1, 2, \ldots, n$$

The result is the sensitive data matrix $D = (d_{ij})_{m \times n}$.

2.2 Privacy Preference Quantification

Generally, sensitivity in structured data tables is classified by impact of information leakage. For instance, "tumor" is more sensitive than "cold" in disease attributes. However, individuals' perception of data sensitivity is fuzzy, and privacy preference reflects unwillingness to disclose personal data.

Definition 4 (Privacy Preference Vector). Let p_{ij} be the weight representing an individual' s unwillingness to disclose sensitive attribute SA_j. The vector $p_i = (p_{i1}, p_{i2}, \ldots, p_{in})$ formed by weights for all sensitive attributes of

one individual is called a privacy preference vector, where $0 \leq p_{ij} \leq 1$ and $\sum_{j=1}^{n} p_{ij} = 1$.

The privacy preference vector can be determined subjectively or inferred from historical data. Its 2-norm $\|p_i\|_2$ reflects privacy preference type: when $\|p_i\|_2$ approaches 1, the individual highly values one or few sensitive attributes while ignoring others; when $\|p_i\|_2$ approaches $\sqrt{1/n}$, the individual shows no privacy preference, valuing all attributes equally.

The privacy preference matrix $P = (p_{ij})_{m\times n}$ comprises all individuals' vectors. Combining sensitive data matrix D with P yields the privacy-preference-weighted sensitive data matrix $G = D \circ P = (d_{ij} \cdot p_{ij})_{m\times n}$.

To describe temporal privacy preference changes and quantify timeliness, three privacy preference function types are introduced.

Definition 5 (Privacy Preference Function). Let $\varphi(t)$ be a function on $[0, +\infty)$ with range $[0, 1]$, called a privacy preference function:

- **Type I:** Individuals increasingly value a sensitive attribute over time with bounded concern: $\varphi(t)$ is a bounded increasing function with $\inf_{t\in[0,+\infty)} \varphi(t) = a$ and $\sup_{t\in[0,+\infty)} \varphi(t) = b$, where $a < b$.
- **Type II:** Individuals increasingly ignore a sensitive attribute over time with bounded disregard: $\varphi(t)$ is a bounded decreasing function.
- **Type III:** Individuals' concern for a sensitive attribute remains unaffected by time: $\varphi(t)$ is constant.

These three functions concisely describe temporal preference changes; more complex patterns can be constructed piecewise from them. Similar to the privacy preference matrix, construct privacy preference function matrix $P(t) = (p_{ij}(t))_{m\times n}$. Combining with sensitive data matrix D yields time-varying privacy-preference-weighted matrix $G(t) = D \circ P(t) = (d_{ij} \cdot p_{ij}(t))_{m\times n}$.

2.3 Privacy Protection Model Analysis

Before data release, privacy protection techniques process the data, falling into two categories: encryption-based and non-encryption-based.

a) **Encryption-based techniques** provide optimal privacy protection—encrypted data reveals no privacy information but becomes unusable. Thus, the sensitive data matrix becomes zero matrix: $D \to D' = 0$.

b) **Non-encryption techniques** include distortion-based and anonymity-based methods. Distortion causes deviation from true data, modeled as matrix perturbation: $D \to D' = D + \Delta$, where $\Delta = (\Delta d_{ij})_{m\times n}$ is the deviation matrix.

Anonymity makes it difficult for attackers to identify individuals, protecting privacy while allowing probabilistic recovery of original sensitive data. From

the matrix perspective, anonymity hides sensitive information with probability q_{ij}, transforming element d_{ij} to 0 with probability $q_{ij} \in [0, 1]$. This change is quantified using expectation:

$$d'_{ij} = d_{ij} \cdot (1 - q_{ij})$$

Through this quantitative analysis, privacy protection degree and post-processing data utility can be evaluated.

2.4 Privacy and Data Utility Measurement

Through numerical processing, privacy preference quantification, and privacy protection model analysis, measuring privacy and utility in structured data reduces to measuring sensitive data matrices, providing intuitive understanding of privacy information.

Let $\mathcal{D}$ be the set of sensitive data matrices from structured data numerical processing (note: "=" here does not mean matrix equality). Define distance $\rho(D_1, D_2) = \|D_1 - D_2\|_F$. Then $(\mathcal{D}, \rho)$ forms a metric space.

Proof: $\mathcal{D}$ is nonempty. For any $D_1, D_2 \in \mathcal{D}$, $\rho(D_1, D_2) = \|D_1 - D_2\|_F \geq 0$, satisfying Definition 1 property (a). For any $D_1, D_2, D_3 \in \mathcal{D}$, $\rho(D_1, D_3) = \|D_1 - D_3\|_F \leq \|D_1 - D_2\|_F + \|D_2 - D_3\|_F = \rho(D_1, D_2) + \rho(D_2, D_3)$, satisfying property (b). Thus $(\mathcal{D}, \rho)$ is a metric space.

The privacy amount in structured data is measured by the "size" of sensitive data matrix points in metric space $\mathcal{D}$. Naturally, norm defines this size.

Definition 6 (Privacy Amount). For $D \in \mathcal{D}$, let $L(D)$ denote the privacy amount of sensitive data matrix D:

$$L(D) = \max\{\|D\|_F\}$$

For privacy-preference-weighted matrix G, $L(G) = \max\{\|G\|_F\}$. For time-varying matrix $G(t)$, $L(G(t)) = \max\{\|G(t)\|_F\}$.

Data utility measurement requires a reference point. Since no specific publishing environment is given, utility is measured by information loss compared to original data: each original data element' s amount is normalized to 1, with processed data amounts in $[0, 1]$.

Definition 7 (Data Utility). Let D be the original sensitive data matrix and D' the processed matrix with identical structure. Their data amounts are represented by $\|D\|_F$ and $\|D'\|_F$. Then utility is:

$$U(D|D') = \frac{\|D'\|_F}{\|D\|_F}$$

where $U(D|D') \in [0,1]$, with values closer to 1 indicating higher fidelity to original data.

Privacy protection degree is another crucial metric, evaluating privacy protection model quality alongside data utility. This paper measures it by privacy amount reduction after protection.

Let M be a privacy protection algorithm transforming D to D'. The privacy protection degree of M on D is:

$$\mathrm{PPD}_M(D) = \frac{L(D) - L(D')}{L(D)}$$

2.5 Model Applicability

Reference [19] comprehensively lists privacy metrics: uncertainty measures, information gain/loss measures, dataset similarity measures, indistinguishability measures, adversary success probability measures, and time/error/precision measures. The proposed model quantifies privacy amount, data utility, and protection degree for structured data in publishing contexts. Its methods apply to information loss measurement, dataset similarity measurement, and beyond —e.g., user identity anonymity and login behavior untraceability in authentication [26,27] can use anonymity-based quantification; access control policies can be measured after structuralization. The model aims to help data owners intuitively grasp privacy information, demonstrated through instance analysis below.

---

3 Instance Analysis

Government data contains extensive personal information with high mining value, but sensitivity concerns prevent publication, often requiring contractual data sharing instead. Privacy and utility metrics enable publishers to effectively control privacy risks and assess leakage. For accuracy and scientific rigor, experiments use the public UCI Machine Learning Repository' s Adult dataset (48,842 records, 14 attributes). The first 5 records with Age, Education, and Occupation attributes form experimental dataset D_1, with simple modifications creating D_2 for comparison (Table 2). Differential privacy protects D_1 to analyze privacy amount, privacy-preference-weighted amount, data utility, and protection degree.

Non-negative Numerical Mapping: Based on sensitivity levels for Age, Education, and Occupation:

- Age: $f_1(x) = \begin{cases} 1 & x \in [0,50] \\ 0 & x > 50 \end{cases}$
- Education: $f_2(\text{Bachelors}) = 0.71$, $f_2(\text{HS-grad}) = 0.50$

- Occupation: f_3(Adm-clerical) = 0.34, f_3(Exec-managerial) = 0.78, f_3(Handlers-cleaners) = 0.95, f_3(Prof-specialty) = 0.65

Numerical processing yields sensitive data matrices:

$$D_1 = \begin{bmatrix} 0.44 & 0.50 & 0.95 \\ 0.44 & 0.50 & 0.95 \\ 0.00 & 0.50 & 0.65 \\ 0.00 & 0.50 & 0.65 \\ 0.48 & 0.71 & 0.34 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0.44 & 0.50 & 0.95 \\ 0.44 & 0.50 & 0.95 \\ 0.00 & 0.50 & 0.65 \\ 0.00 & 0.50 & 0.34 \\ 0.88 & 0.50 & 0.78 \end{bmatrix}$$

Privacy Amount: $L(D_1) = 2.3223$, $L(D_2) = 2.3944$. D_2 shows larger privacy amount, consistent with Table 2' s principle that rarer occupations increase privacy.

Privacy Preference Functions: For dataset D_1, the privacy preference function matrix is:

$$P(t) = \begin{bmatrix} \frac{0.33t}{1+0.33t} & \frac{0.33t}{1+0.33t} & \frac{0.34t}{1+0.34t} \\ \frac{0.0051t}{1+0.0051t} & \frac{0.005t}{1+0.005t} & \frac{0.33t}{1+0.33t} \\ \frac{0.004t}{1+0.004t} & \frac{0.33t}{1+0.33t} & \frac{0.34t}{1+0.34t} \\ \frac{0.003t}{1+0.003t} & \frac{0.0011t}{1+0.0011t} & \frac{0.001t}{1+0.001t} \\ \frac{0.001t}{1+0.001t} & \frac{0.33t}{1+0.33t} & \frac{0.34t}{1+0.34t} \\ \frac{0.003t}{1+0.003t} & \frac{0.006t}{1+0.006t} & \frac{0.006t}{1+0.006t} \\ \frac{0.33t}{1+0.33t} & \frac{0.33t}{1+0.33t} & \frac{0.34t}{1+0.34t} \\ \frac{0.006t}{1+0.006t} & \frac{0.006t}{1+0.006t} & \frac{0.006t}{1+0.006t} \end{bmatrix}$$

The time-varying privacy-preference-weighted matrix is $D_1(t) = D_1 \circ P(t)$. Figure 1 [Figure 1: see original paper] shows $L(D_1(t))$ initially increasing then decreasing, stabilizing near $t = 1.8$, indicating reduced concern for some sensitive attributes over time.

Differential Privacy Protection: Generate Laplace noise using inverse CDF and uniform random sequences. Let a be uniform on $[0, 1]$, $\mu = 0$, $\sigma = 0.01$. Laplace random number δ with standard deviation σ is:

$$\delta = \mu - \sigma \cdot \text{sgn}(a - 0.5) \cdot \ln(1 - 2|a - 0.5|)$$

Noise matrix Δ is:

$$\Delta = \begin{bmatrix} 0.0052 & 0.0296 & 0.0068 \\ 0.0329 & 0.0066 & 0.0024 \\ 0.0070 & 0.0012 & 0.0136 \\ 0.0058 & 0.0043 & 0.0014 \\ 0.0066 & 0.0000 & 0.0121 \end{bmatrix}$$

Protected matrix $D_1' = D_1 - \Delta$:

$$D_1' = \begin{bmatrix} 0.4348 & 0.4704 & 0.9432 \\ 0.0000 & 0.4934 & 0.6476 \\ 0.4730 & 0.7088 & 0.3264 \\ 0.0000 & 0.4957 & 0.3386 \\ 0.8734 & 0.5000 & 0.7679 \end{bmatrix}$$

Results: $L(D_1') = 2.3127$, $U(D_1|D_1') = 0.9877$, $\mathrm{PPD}_M(D_1) = 0.0041$. With small σ, protection degree is low while utility remains high. Increasing noise raises protection degree but reduces utility, as shown in Figure 2 [Figure 2: see original paper].

The instance analysis demonstrates that the proposed model effectively reflects privacy amount, protection degree, and utility changes, providing quantitative basis for assessing data publication risks.

---

4 Conclusion

No complete privacy metric theory currently exists. Common methods based on probability statistics, information theory, and set pair analysis have limitations. This paper proposes a privacy and data utility metric model for structured data, constructing metric spaces to measure information distances and defining privacy amount. To quantify privacy and utility: (1) non-numeric data is transformed into computable matrices; (2) three privacy preference functions capture subjective, time-varying sensitivity; (3) privacy protection model impacts are quantified; (4) matrix distances measure privacy and utility. Instance analysis confirms the model effectively reflects privacy amount and utility changes, serving as a quantitative method for the privacy-utility tradeoff.

References

[1] Li Ninghui, Li Tiancheng, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. Proc of the 23rd International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2007.

[2] Li Ninghui, Li Tiancheng, Venkatasubramanian S. Closeness: a new privacy measure for data publishing. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(7): 943-956.

[3] Zhang Jianpei, Xie Jing, Yang Jing, et al. A t-closeness privacy model based on sensitive attribute values semantics bucketization. Journal of Computer Research and Development, 2014, 51(1): 126-137.

[4] Gkountouna O, Terrovitis M. Anonymizing collections of tree-structured data. IEEE Trans on Knowledge & Data Engineering, 2015, 27(8): 2021-2034.

[5] Yuji Y, Kouichi I. k-presence-secrecy: practical privacy model as extension of k-anonymity. IEICE Trans. on Information& System, 2017(4): 730-740.

[6] Li Xiangyang, Zhang Chunhong, Jung T, et al. Graph-based privacy-preserving data publication. Proc of the 35th Annual IEEE International Conference on Computer Communications. Piscataway, NJ: IEEE Press, 2016: 1-9.

[7] Shannon C E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27(3): 379-423.

[8] Díaz C, Seys S, Claessens J, et al. Towards measuring anonymity. Proc of the 2nd International Conference on Privacy Enhancing Technologies. Berlin: Springer, 2002: 54-68.

[9] Clauß S, Stefan S. Structuring anonymity metrics. Proc of the 2nd ACM Workshop on Digital Identity Management. New York: ACM Press, 2006: 55-62.

[10] Hoh B, Gruteser M, Xiong Hui, et al. Preserving privacy in gps traces via uncertainty-aware path cloaking. Proc of the 14th ACM conference on Computer and communications security. New York: ACM Press, 2007: 161-171.

[11] Ma Zhendong, Kargl K, Weber M. A location privacy metric for V2X communication systems. Proc of IEEE Sarnoff Symposium. Piscataway, NJ: IEEE Press, 2009: 1-6.

[12] Ma Zhendong, Kargl K, Weber M. Measuring location privacy in V2X communication systems with accumulated information. Proc of the 6th IEEE International Conference on Mobile Adhoc and Sensor Systems. Piscataway, NJ: IEEE Press, 2009: 322-331.

[13] Shokri R, Freudiger J, Jadliwala M, et al. A distortion-based metric for location privacy. Proc of the 8th ACM Workshop on Privacy in the Electronic Society. New York: ACM Press, 2009: 21-30.

[14] Chen Xihui, Pang Jun. Measuring query privacy in location-based services. Proc of the 2nd ACM Conference on Data and Application Security and Privacy. New York: ACM Press, 2012: 49-60.

[15] Yang Yuhao, Lutes J, Li Fengjun, et al. Stalking online: On user privacy in social networks. Proc of the 2nd ACM Conference on Data and Application Security and Privacy. New York: ACM Press, 2012: 37-48.

[16] Peng Changgen, Ding Hongfa, Zhu Yijie, et al. Information entropy models and privacy metrics methods for privacy protection. Journal of Software, 2016, 27(8): 1891-1903.

[17] Zhao Keqin. Set pair analysis and its preliminary application. Hangzhou: Zhejiang Science and Technology Press, 2000.

[18] Yan Yan, Hao Xiaohong, Wang Wanjun. A set pair analysis method for privacy metric. Engineering Journal of Wuhan University, 2015, 48(6): 883-890.

[19] Wagner I, Eckhoff D. Technical privacy metrics: a systematic survey. ACM Computing Surveys, 2018, 51(3): articleNo 57.

[20] Wang Lingling, Ma Chunguang, Liu Guozhu. Survey on metrics for location-based privacy protection mechanisms. Application Research of Computers, 2017, 34(3): 647-652.

[21] Xiong Jinbo, Wang Minshen, Tian Youliang, et al. Research progress on privacy measurement for cloud data. Journal of Software, 2018, 29(7): 1963-1980.

[22] Wang Lu, Meng Xiaofeng. Location privacy preservation in big data era: a survey. Journal of Software, 2014, 25(4): 693-712.

[23] Zhang Xuejun, Gui Xiaolin, Wu Zhongdong. Privacy preservation for location-based services: a survey. Journal of Software, 2015, 26(9): 2373-2395.

[24] Xia Daoxing. Real variable function theory and functional analysis, volume 2. Beijing: Higher Education Press, 2010: 1-107.

[25] Dwork C, Roth A. The algorithmic foundations of differential privacy. Boston: Now Publishers Inc., 2014.

[26] Wang Ding, Wang Ping. On the anonymity of two-factor authentication schemes for wireless sensor networks: attacks, principle and solutions. Computer Networks, 2014, 73(11): 41-57.

[27] Wang Ding, Wang Ping. Two birds with one stone: two-factor authentication with security beyond conventional bound. IEEE Trans on Dependable and Secure Computing, 2018, 15(4): 708-722.

[28] Dasgupta B, Mobasheri N, Yero I G. On analyzing and evaluating privacy measures for social networks under active attack. Information Sciences, 2019, 473(1): 87-100.

[29] Ahmad A, Mukkamala R. A novel information privacy metric. Proc of the 14th International Conference on Information Technology. Cham: Springer, 2018: 221-226.

[30] Zhao Yuchen, Wagner I. POSTER: evaluating privacy metrics for graph anonymization and de-anonymization. Proc of Asia Conference on Computer and Communications Security. New York: ACM Press, 2018.

Note: Figure translations are in progress. See original paper for figures.