2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th Intl. Conference on Data Science and Systems

Non-Asymptotic Bound on the Performance of k-Anonymity against Inference Attacks

Ping Zhao[1,2] Hongbo Jiang[3] Chen Wang[1] Haojun Huang[4]
[1]Huazhong University of Science and Technology, Wuhan 430074 China
[2]Donghua University, Shanghai 201620 China
[3]Hunan University, Changsha 410082 China
[4]China University of Geosciences, Wuhan 430074 China

***Abstract*—Internet of Things (IoT) applications bring in a great convenience for human's life, but users' data privacy concern is the major barrier towards the development of IoT. k-Anonymity is a method to protect users' data privacy, but it is presently known to suffer from inference attacks. Thus far, existing work only relies on a number of experimental examples to validate k-anonymity's performance against inference attacks, and thereby lacks of a theoretical guarantee. To tackle this issue, in this paper we propose the first theoretical foundation that gives a non-asymptotic bound on the performance of k-anonymity against inference attacks, taking into consideration of adversaries' background information. The main idea is to first quantify adversaries' background information, and from the point of the view of adversaries, classify users' data into four kinds: independent with unknown data values, local dependent with unknown data values, independent with certain known data values, and local dependent with certain known data values. We then move one step further, theoretically proving the bound on the performance of k-anonymity corresponding to each of the four kinds of users' data through cooperating with the noiseless privacy. We argue that such a theoretical foundation links k-anonymity with noiseless privacy, theoretically proving k-anonymity provides noiseless privacy. Additionally, our work theoretically explains why k-anonymity is vulnerable to inference attacks using the modified Stein method. We believe that our work can bridge the gap between design and evaluation, enabling a designer to construct a more practical k-anonymity technique in real-life scenarios to resist inference attacks.**

***Index Terms*—k-Anonymity, noiseless privacy, inference attacks, non-asymptotic bound.**

I. Introduction

In the era of Internet of Things (IoT), an increasing number of devices embedded with electronics, software, sensors, actuators, are connected via the Internet, enabling various IoT applications, e.g., smart grid, smart cities, smart transportation system, etc., that offer a great benefit for human's life [1], [2], [3], [4], [5], [6]. However, such convenience does not come for free, as users' data created by these IoT devices are collected in these IoT applications. As a result, users' data privacy may be disclosed to the untrusted data aggregator, and more seriously, more sensitive personal information implied in users' data such as one's social relationship and political beliefs can be invaded.

k-Anonymity is proposed to protect users' data privacy in such a scenario, which blurs a user's data into a cloaked data set such that the user's data cannot be distinguished from at least $(k-1)$ data of other users in this set [7]. However, k-anonymity is susceptible to inference attacks where adversaries identify every user's data utilizing the background information about users' data, and thus cannot fully protect users' privacy in IoT [8], [9], [10], [11].

To protect users' data privacy against inference attacks, a bunch of work has focused on refining the data cloaked in the same set via inserting or removing edges (or nodes) into or from the cloaked data [12], [13], [14], cloaking users' data with same attributes and structural information in a set [15] [16], generalizing users' data in terms of structural features [17], [18], or considering a wide range of users' data and quantifying both the privacy and utility measurements of the cloaked data [19]. However, all the work has no rigorous theoretical analysis of the performance of the proposed techniques, only relying on a number of experimental examples to validate the performance against inference attacks. The question then naturally rises as what the performance of k-anonymity exactly is when suffering from inference attacks.

In this paper, we propose the first theoretical foundation that gives a non-asymptotic bound on the performance of k-anonymity against inference attacks, taking into consideration of adversaries' background information. To the best of our knowledge, it is the first work that gives the non-asymptotic bound on the performance of k-anonymity against inference attacks. In addition, it offers several salient features. First, it is the first work to link k-anonymity with noiseless privacy, and theoretically prove "cloaking in a set of size no less than k" offers noiseless privacy. Second, it thoroughly and theoretically analyse why k-anonymity is susceptible to inference attacks employing the modified Stein method. Third, it is the first work to fill up the gap between design and evaluation, enabling a designer to propose a more practical privacy preserving k-anonymity technique in real-life scenarios to resist inference attacks. In summary, we believe it is a step towards more practical constructions of k-anonymity.

II. Preliminary

A. Syntactic Sensitivity

Definition 1: The syntactic sensitivity s of k-anonymity $\mathcal{F}$ with respect to the input data set $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$

This work was supported in part by the National Natural Science Foundation of China under Grants 61502192, 61572219, 61502193, 61671216, 61702204 and 41701479; by the China Postdoctoral Science Foundation under Grants 2017T100556; and by the Fundamental Research Funds for the Central Universities under Grant 2016YXMS297 and 2016JCTD118. The corresponding author of this paper is Chen Wang.

978-1-5386-6614-2/18/$31.00 ©2018 IEEE
DOI 10.1109/HPCC/SmartCity/DSS.2018.00106

is [20],

$$s = \max \mid Vol(\mathcal{F}(\mathcal{D})) - Vol(\mathcal{F}(\mathcal{D}^*)) \mid, \qquad (1)$$

where $\mathcal{D}^*$ is any a data set with one data is removed from or added to $\mathcal{D}$ (hereafter $\mathcal{D}^*$ is called adjacent data of $\mathcal{D}$); $\mathcal{F}(\mathcal{D})$ is the output of $\mathcal{F}$ with respect to $\mathcal{D}$; and $Vol(\mathcal{F}(\mathcal{D}))$ is the number of users' data cloaked in one set by k-anonymity $\mathcal{F}$.

For instance, assume $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ is n users' friendship data, $\mathcal{D}_i$ is user u_i's friendship data, and each tuple in $\mathcal{D}_i$ is the digital label (i.e., the pseudo-identity) of one of u_i's friends (e.g., 1 in $\mathcal{D}_i = \{1, 2, 3, \ldots\}$). Denote the set of adjacent data of $\mathcal{D}$ as $\Omega = \{\mathcal{D}_1^*, \mathcal{D}_2^*, \ldots, \mathcal{D}_n^*\}$, where $\mathcal{D}_i^* = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{i-1}, \mathcal{D}_{i+1}, \ldots \mathcal{D}_n\}$ is a adjacent data of $\mathcal{D}$. Assume $\mathcal{F}(\mathcal{D}) = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k\}$, namely friendship data of u_1, …, u_k is cloaked in one dataset. For example, if $\mathcal{F}$ cloaks $(k+1)$ users' friendship data in one dataset with respect to any a adjacent data of $\mathcal{D}$, $s = 1$. Also, if $\mathcal{F}$ cloaks $(k+3)$ users' friendship data in one dataset when input is a specific adjacent data $\mathcal{D}_i^*$, and it cloaks k users' friendship data in one dataset with respect to each of the remaining adjacent data, according to Definition 1, $s = 3$.

B. Adversary Model

We consider users' data cloaked by a specific k-anonymity technique are collected by a data aggregator in IoT applications, e.g., data aggregation, crowdsensing, etc. The data aggregator may be untrusted and interested in identifying the data of each user through launching inference attacks, using its background information. On the other hand, the aggregator may also disclose the cloaked data, for some commercial interests, to researchers or industry who may be malicious and conduct inference attacks to distinguish users' data. Note that hereafter, untrusted data aggregator, malicious researchers, and industry, etc. are called adversaries, for the ease of description.

Adversaries can obtained background information through various means: (1) Many companies share users' data, e.g., Twitter shares data with its partner IBM [21]; (2) A large number of data in e.g., Twitter [22], YouTube [22] can be crawled; (3) Users' data is widely available on e.g., SNAP, CMU Datasets [23], etc. These available background information enables adversaries to get the access to the exact value of users' data or the dependencies among users' data. For example, in Fig. **??**, the background information, ambulatory care data of Allen enables adversaries to identify Allen's traces (i.e., *exact value*); in Figs. **??** and **??** adversaries can infer the *dependencies* among cloaked locations employing the structure similarity between the cloaked locations and the background information (i.e., friendships in social networks). Moreover, the *syntactic sensitivity* s can also be disclosed to adversaries, since the outputs of k-anonymity techniques are known to the untrusted data aggregator, and it thus can compute the syntactic sensitivity s through observing the number of users' data cloaked in one set.

C. Quantification of Adversaries' Background Information

We quantify adversaries' background information as follows: (a) the syntactic sensitivity s; (b) the dependencies among λ_1 percents of users' data, and the syntactic sensitivity s ($\lambda_1 < 1$); (c) the exact value of λ_2 percents of users' data, and the syntactic sensitivity s ($\lambda_2 < 1$); (d) the exact value of λ_2 percents of users' data, the dependencies among λ_1 percents of users' data, and the syntactic sensitivity s. For the ease of description, we denote Level-I, II, III, IV inference attacks corresponding to the above four kinds of background information.

On the basis of the qualified adversaries' background information, users' data can be classified into the following four kinds, from the perspective of adversaries:

- Independent data $\mathcal{D}$ with known syntactic sensitivity s;
- Local dependent data $\mathcal{D}$ with disclosed syntactic sensitivity s and dependencies of λ_1 percent of data.
- Independent data $\mathcal{D}$ with known syntactic sensitivity s and a subset Θ ($\Theta \subseteq \mathcal{D}$) containing λ_2 percent of data;
- Local dependent data $\mathcal{D}$ with a known subset Θ ($\Theta \subseteq \mathcal{D}$) containing λ_2 percent of data, the disclosed syntactic sensitivity s, and the dependencies of λ_1 percent of data.

D. Noiseless Privacy

Noiseless privacy makes efforts toward answering the question "Is it always necessary to add noise to the output to achieve provable privacy guarantees?" [24], and provides an alternative approach to achieve privacy preservation. It exploits the inherently uncertainty in the database without injecting noise into the output. Its formal definition is as follows:

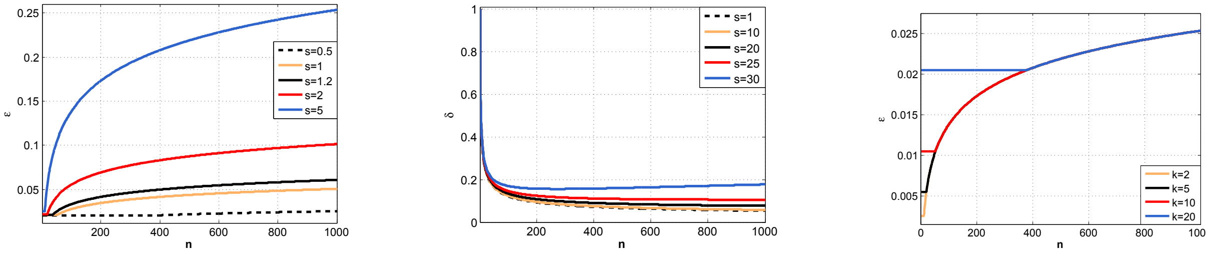
Definition 2: Denote $\mathcal{F}$: $\mathcal{D}^n \to \mathcal{Y}$ a privacy mechanism. $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy, if for all $\mathcal{O} \in \mathcal{Y}$ and all $X \in \mathcal{D}$, $X^* \in \mathcal{D}$ (X^* is obtained by removing or adding a tuple from or to X), the following holds [24]:

$$Pr[\mathcal{F}(X) \in \mathcal{O}] \leq \exp(\epsilon) Pr[\mathcal{F}(X^*) \in \mathcal{O}] + \delta. \qquad (2)$$

It implies that the input $\mathcal{D}^n$ inherently results in the uncertainty of adversaries and thus can be protected by such uncertainty instead of the randomness of mechanisms. So even deterministic mechanisms can guarantee data privacy, satisfying noiseless privacy without adding external noise. Motivated by Definition 2, we concentrate on theoretically proving the privacy guarantees that the deterministic mechanism k-anonymity provides.

E. Theoreitcal Basis

We first consider an impractical case where adversaries do not have any background information about the cloaked data. To be more concrete, adversaries do not know the values of users' data, the dependencies among users' data, and even the syntactic sensitivity. In this case, from the respective of adversaries, any a k-anonymity technique $\mathcal{F}$ works as follows: randomly selecting no less than k tuples from the set of users' data $\mathcal{D}$, and cloaking these tuples into one cloaked set.



(a) (b) (c)

Fig. 1. (a) The parameter ϵ varies with syntactic sensitivity s, where $k = 10$, $\epsilon_1 = 0.1$, $\sqrt{\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)} = 50$. (b) The parameter δ varies with syntactic sensitivity s, where $\sqrt{\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)} = 50$, $\sum_{i=1}^{n}(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3) = 1$, $\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}} = 100$. (c) The parameter ϵ varies with privacy parameter k, where $s = 1$, $\epsilon_1 = 0.1$, $\sqrt{\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)} = 50$.

Furthermore, we theoretically prove the performance of the k-anonymity technique $\mathcal{F}$ via cooperating with noiseless privacy as follows.

Theorem 1: Denote $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$. $\mathcal{F}$ randomly selects α ($\alpha \geq k$) tuples in $\mathcal{D}$ with probability p. (I) For all $\delta \geq \frac{n!p^k q^{n-k}}{k!(n-k)!} + q^n$, $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy with $\epsilon = \max\{\ln((\frac{n+1}{np - \sqrt{-\frac{1}{2}n\ln[\delta + \exp(\frac{-2(np-k+1)^2}{n})]}} - 1)\frac{p}{1-p}), \ln((\frac{n+1}{n - np - \sqrt{-\frac{1}{2}n\ln[\delta + \exp(\frac{-2(np-k+1)^2}{n})]}} - 1)\frac{1-p}{p})\}$ (II) For all $\epsilon > 0$, $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy where $\delta = \max\{\exp[\frac{-2(np - \frac{p(n+1)}{\exp(\epsilon)(1-p)+p})^2}{n}] - \exp[\frac{-2(np-(k-1))^2}{n}], \exp[\frac{-2(\frac{\exp(\epsilon)np-1+p}{\exp(\epsilon)p+1-p})^2}{n}]\}$.

Proof: See Appendix. A. ■

Theorem 1 gives an explicitly bound on the performance of k-anonymity when adversaries do not have any background information. However, Theorem 1 only considers a trivial scenario where the adversaries do not have any background information about the cloaked data. In practical scenario, adversaries can obtained background information through various means, thereby deteriorating the performance of k-anonymity techniques. Therefore, in the following, we consider strategic adversaries that get background information about the cloaked data, and bound the performance of k-anonymity techniques against such adversaries on the basis of Theorem 1.

III. Performance Bounds Analysis

As regard to the performance of the various k-anonymity techniques against inference attacks, we propose to bound their performance utilizing the background information of adversaries. Such a proposal stems from the observation in Theorem 1 that the sophisticatedly selected data will be the ones that are randomly selected when adversaries do not have any background information about these selected data. That is, adversarial uncertainty about the cloked data can be leveraged to protect users' data privacy.

A. Bounds on Performance against Level-I Inference Attacks

In this part, we consider any a k-anonymity technique $\mathcal{F}$ that selects no less than k users' data from the dataset $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ and cloak these selected data into a set. Since only the syntactic sensitivity is disclosed to adversaries, the cloaked data (i.e., the output of the k-anonymity technique $\mathcal{F}$) is independent with unknown values. In such a case, motivated by Theorem 1, we theoretically prove the performance of $\mathcal{F}$ using noiseless privacy as follows.

Lemma 2: For variable $Q \sim \mathcal{N}(0, \sigma^2)$, $\sigma \geq \frac{cs}{\epsilon_1}$, and c meets

$$\begin{cases} \ln\frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln\frac{1}{\sqrt{2\pi}\delta_1}, \\ \frac{c^2 s}{\epsilon_1} - \frac{s}{2} > k, \end{cases}$$

we get

$$Pr[u + Q \in \mathcal{O}] \leq \exp(\epsilon_1)Pr[v + Q \in \mathcal{O}] + \delta_1,$$

where S is the output of k-anonymity based technique $\mathcal{F}$.

Proof: See Appendix. B ■

Lemma 2 extends the theory of differential privacy in [20], by considering the constraints of k-anonymity techniques (i.e., more than k tuples) and giving a tighter bound on privacy parameters (ϵ_1, δ_1).

Lemma 3: Denote variables $X = \{X_1, \ldots, X_n\}$ with variances $\sigma_1^2, \ldots, \sigma_n^2$ and finite third absolute moments $\rho_1, \ldots, \rho_n$. For the normed sum function S_n of X and the standard normal random variable N, the following holds [25]:

$$\sup_{x \in R} \mid Pr[S_n \leq x] - Pr[N \leq x] \mid \leq \frac{\varpi \sum_{i=1}^{n} \rho_i^2}{\sum_{i=1}^{n} \sigma_i^2},$$

where $\varpi \leq 0.5591$.

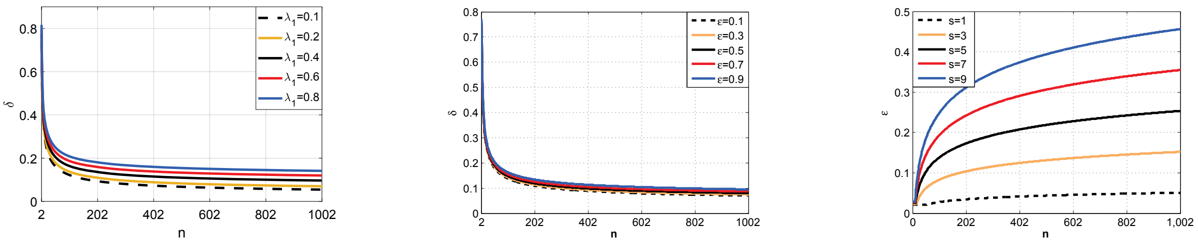
On the basis of Lemmas 2 and 3, we theoretically analyse the performance of k-anonymity technique $\mathcal{F}$ as follows.

Theorem 4: The k-anonymity technique $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy with $\delta = \frac{1.1182\sum_{i=1}^{n}(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}}(1 + \exp(\sqrt{\frac{s^2\ln n}{\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)}})) + \frac{1}{\sqrt{n}}$ and $\epsilon = \frac{c's}{\sqrt{\sum_{i=1}^{n}(\varphi_i^2 - \pi_i^2)}}$, where $mean(\mathcal{D}_i) = \pi_i$, $mean(\mathcal{D}_i^2) = \varphi_i^2$, $mean(\mathcal{D}_i^3) = \chi_i^3$, and c' is the minimum value of c that meets:

$$\begin{cases} \ln\frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln\frac{\sqrt{n}}{\sqrt{2\pi}}, \\ \frac{c^2 s}{\epsilon_1} - \frac{s}{2} > k, \end{cases}$$

Proof: See Appendix. C. ■

Theorem 4 enables a designer to construct a practical and optimal k-anonymity technique in real-life scenarios to resist



(a) (b) (c)

Fig. 2. (a) The privacy parameter δ varies with λ_1, where $\sigma = 50$, $\sum_{i=1}^{n} \chi_i^3 = 5$, $\sqrt{\sum_{i=1}^{n} \xi_i^4} = 1$, and $\epsilon = 0.1$. (b) The privacy parameter δ varies with ϵ, where $\sigma = 50$, $\sum_{i=1}^{n} \chi_i^3 = 5$, $\sqrt{\sum_{i=1}^{n} \xi_i^4} = 1$, and $\lambda_1 = 0.2$. (c) The privacy parameter ϵ varies with s, where $\varepsilon_1 = 0.1$ and $k = 10$.

inference attacks, considering the background information of adversaries. As shown in Figs. 1(a) and 1(b), the privacy parameters ϵ and δ increase with the syntactic sensitivity s of the designed k-anonymity technique $\mathcal{F}$. That is because the output of the proposed $\mathcal{F}$ with a larger syntactic sensitivity s is sharply changed, thus resulting in a larger probability that breaches Formula (2). Moreover, it can be observed in Figs. 1(a) and 1(b), privacy parameter δ is more robust to the syntactic sensitivity s than ϵ. In addition, in Fig. 1(c), the performance of the designed k-anonymity technique $\mathcal{F}$ is affected by the privacy requirement k of users. That is not surprising, as more users' data cloaked in a set results in more generalization of the cloaked data, i.e., a larger ϵ.

B. Bounds on Performance against Level-II Inference Attacks

In Level-II attacks, the dependence among λ_1 percents of users' data is disclosed to adversaries, except for the syntactic sensitivity s, which enables adversaries to launch more offensive inference attacks, e.g., de-anonymity attacks. Accordingly, the performance of the k-anonymity based technique $\mathcal{F}$ follows:

Theorem 5: The k-anonymity based technique $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy with $\epsilon = \frac{c's}{\sigma}$ and $\delta = 2(\frac{2}{\pi})^{1/4}\sqrt{\frac{\lambda_1^2}{\sigma^3}\sum_{i=1}^{n}\chi_i^3 + \frac{\lambda_1^{\frac{3}{2}}\sqrt{26}}{\sigma^2\sqrt{\pi}}\sqrt{\sum_{i=1}^{n}\xi_i^4}}(1+\exp(\epsilon))+\frac{1}{\sqrt{n}}$, where $\sigma^2 = var[\sum_{i=1}^{n}\mathcal{D}_i]$, $mean(\mathcal{D}_i^3) = \chi_i^3$, $mean(\mathcal{D}_i^4) = \xi_i^4$, $\lambda_1 = \max_{1\leq i\leq n} \mid N_i \mid$, N_i $(i = 1, \ldots, n)$ are dependent neighborhoods, and c' is the minimum value of c satisfying:

$$\begin{cases} \ln\frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln\frac{\sqrt{n}}{\sqrt{2\pi}}, \\ \frac{c^2 s}{\epsilon_1} - \frac{s}{2} > k, \end{cases}$$

Proof: See Appendix. D. ■

Theorem 5 theoretically bounds the performance of the k-anonymity based technique $\mathcal{F}$ (cf. Formulas (33) and (36)). Examples of the bounds, i.e., privacy parameters ϵ and δ are shown in Fig. 2. It can be observed in Fig. 2(a), δ increases with the amount of adversaries' background information, as adversaries can disclose more users' data when they have more background information, thereby resulting in a larger δ. In addition, as shown in Fig. 2(b), δ decreases with the decreasing ϵ, since it is more likely for adversaries to identify a specific user's data through observing the output when ϵ has a larger value. Furthermore, ϵ increases with the syntactic sensitivity s, as the output of $\mathcal{F}$ is more significantly changed with respect to the same input when $\mathcal{F}$ has a larger syntactic sensitivity s.

On the other hand, Theorem 5 motivates the designer to select quantified users' data to lower the bound of the performance of the k-anonymity based technique $\mathcal{F}$ to be proposed. To concrete, the designer can decrease numerator and increase denominator of the formulas that formalize privacy parameters ϵ and δ, through selecting quantified users' data. As such, more users' data privacy can be protected against inference attacks.

C. Bounds on Performance against Level-III Inference Attacks

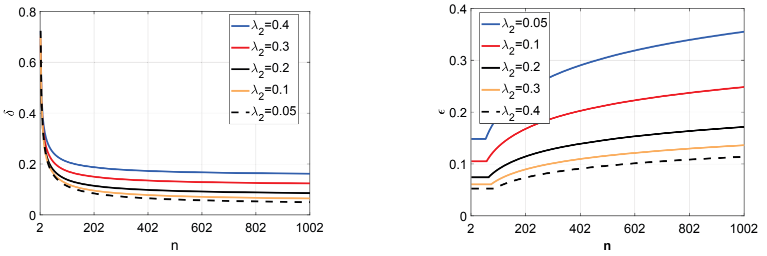
In Level-III attacks, λ_2 percents of data and syntactic sensitivity s are known to adversaries. As analysed in Level-I attacks, we consider any a k-anonymity technique $\mathcal{F}$ that selects no less than k users' data from the dataset $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$ and cloaks these selected data into a set. In such a scenario, the remaining $(1 - \lambda_2)$ percents of data is randomized in adversaries' perspective, and such randomness can effectively hide other data values. Therefore we have

Theorem 6: The k-anonymity based technique $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy with $\delta = \frac{1.1182\sum_{i=1,i\notin\Gamma}^{n}(\chi_i^3-3\varphi_i^2\pi_i+2\pi_i^3)}{\sum_{i=1,i\notin\Gamma}^{n}(\varphi_i^2-\pi_i^2)^{\frac{3}{2}}}(1+\exp(\sqrt{\frac{s^2\ln(n(1-\lambda_2))}{\sum_{i=1,i\notin\Gamma}^{n}(\varphi_i^2-\pi_i^2)}}))+\frac{1}{\sqrt{n(1-\lambda_2)}}$ and $\epsilon = \frac{c's}{\sqrt{\sum_{i=1,i\notin\Gamma}^{n}(\varphi_i^2-\pi_i^2)}}$, where $mean(\mathcal{D}_i) = \pi_i$, $mean(\mathcal{D}_i^2) = \varphi_i^2$, $mean(\mathcal{D}_i^3) = \chi_i^3$, Γ is the set of the indexes of the $n\lambda_2$ users' data that is known to adversaries, and c' is the minimum value c satisfying:

$$\begin{cases} \ln\frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln\frac{\sqrt{n(1-\lambda_2)}}{\sqrt{2\pi}}, \\ \frac{c^2 s}{\epsilon_1} - \frac{s}{2} > k, \end{cases}$$

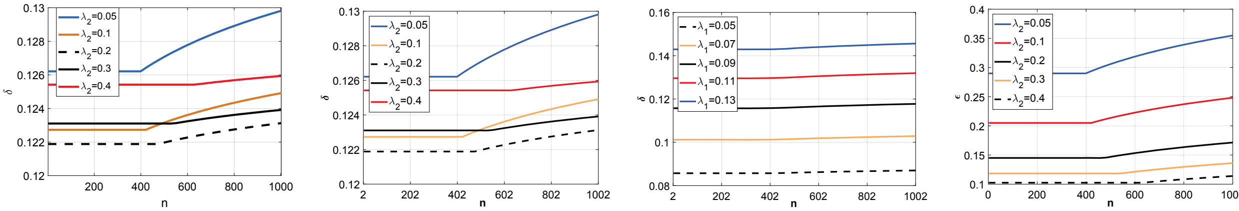
The proof of Theorem 6 is similar to that of Theorem 4, but only differ in the number of data that is not disclosed to adversaries. That is, in the proof of Theorem 6, $n(1 - \lambda_2)$ data contributes to the uncertainness of adversaries while in the proof of Theorem 4, the n data brings in the privacy preservation.

Theorem 6 qualities the deterioration of the k-anonymity technique $\mathcal{F}$ cased by adversaries' background information in Level-III inference attacks (cf. Fig. 3). It can be observed in



(a) (b)

Fig. 3. (a) The privacy parameter δ varies with λ_2, where $s = 1$, $\sum_{i=1,i\notin\Gamma}^{n}(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3) = 10\lambda_2$, $\sum_{i=1,i\notin\Gamma}^{n}(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}} = 100\lambda_2$, $\sum_{i=1,i\notin\Gamma}^{n}(\varphi_i^2 - \pi_i^2) = 10\lambda_2$. (b) The privacy parameter ϵ varies with λ_2, where $\varepsilon_1 = 0.1$, $k = 10$, $\sum_{i=1,i\notin\Gamma}^{n}(\varphi_i^2 - \pi_i^2) = 10\lambda_2$, $s = 1$.



(a) (b) (c) (d)

Fig. 4. (a) (b) The privacy parameter δ varies with λ_2 and λ_1, where $\sum_{i=1,i\notin\Gamma}^{n}\chi_i^3 = 5\lambda_2$, $\sqrt{\sum_{i=1,i\notin\Gamma}^{n}\xi_i^4} = 5\lambda_2$, $\sigma^2 = 1000\lambda_2$, $k = 20$, $s = 1$, $\epsilon_1 = 0.1$, and $\lambda_1 = 0.1$ in (a) ($\lambda_2 = 0.1$ in (b)). (c) The privacy parameter ϵ varies with λ_2, with $k = 20$, $\sigma^2 = 1000\lambda_2$, $s = 1$, and $\epsilon_1 = 0.1$.

Fig. 3(a), privacy parameter δ increases with the amount of adversaries' background information λ_2. That is not surprising, as more background information enables adversaries to disclose more users' data, thereby increasing δ. As shown in Fig. 3(b), privacy parameter ϵ decreases with the amount of adversaries' background information λ_2. The reason is that more background information leads to a larger probability that Formula (2) does not hold, and thus the privacy parameter ϵ is decreased.

Theorem 6 also inspires the designer to propose a practical k-anonymity based privacy preserving technique that achieves the best performance against inference attacks, given the adversaries' background information λ_2 and and s. To be more concrete, the designer can deliberately select these users' data to be cloaked in a set to decrease the numerator of the formula of parameter δ and increase the denominator of the formulas of both δ and ϵ in Theorem 6. The decreased parameter δ means less probability that users' data privacy is disclosed, and the decreased parameter ϵ means that it is more difficult for adversaries to infer users' data privacy through observing the output of the proposed k-anonymity technique $\mathcal{F}$. As a result, more users' data privacy is protected.

D. Bounds on Performance against Level-IV Inference Attacks

In this section, we consider stronger adversaries that have the prior knowledge of the exact values of $n\lambda_2$ users' data, the dependence among $n\lambda_1$ users' data, and the syntactic sensitivity s. The performance of the proposed privacy preserving technique $\mathcal{F}$ against such adversaries is presented as:

Theorem 7: The k-anonymity based technique $\mathcal{F}$ meets (ϵ, δ)-noiseless privacy with $\epsilon = \frac{c's}{\sigma}$ and $\delta = 2(\frac{2}{\pi})^{1/4}\sqrt{\frac{\lambda_1^2}{\sigma^3}\sum_{i=1,i\notin\Gamma}^{n}\chi_i^3 + \frac{\lambda_1^{\frac{3}{2}}\sqrt{26}}{\sigma^2\sqrt{\pi}}\sqrt{\sum_{i=1,i\notin\Gamma}^{n}\xi_i^4}(1 + \exp(\epsilon))} + \frac{1}{\sqrt{n(1-\lambda_2)}}$, where $\sigma^2 = var[\sum_{i=1,i\notin\Gamma}^{n}\mathcal{D}_i]$, $mean(\mathcal{D}_i^3) = \chi_i^3$, $mean(\mathcal{D}_i^4) = \xi_i^4$, $\lambda_1 = \max_{1\leq i\leq n} \mid N_i \mid$, N_i $(i = 1, \ldots, n)$ are the dependent neighborhoods, and c' is the minimum value c that meets:

$$\begin{cases} \ln\frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln\frac{\sqrt{n(1-\lambda_2)}}{\sqrt{2\pi}}, \\ \frac{c^2 s}{\epsilon_1} - \frac{s}{2} > k, \end{cases}$$

The proof of Theorem 7 is analogous to the proof of Theorem 5, only differing in the number of users' data that contributes to the uncertainness of adversaries.

Examples of the bound on the performance of $\mathcal{F}$ are shown in Fig. 4. The privacy parameter δ decreases with the increasing λ_2 when $\lambda_2 \leq 0.2$, and increases with increasing λ_2 when $\lambda_2 \geq 0.3$, as shown in Fig. 4(b). That is because, δ is affected by both λ_2 and ϵ, and ϵ is affected by λ_2 at the same time. Moreover, the background information λ_2 contributes to the probability that adversaries successfully disclose users' data privacy. In addition, Fig. 4(c) shows that δ increases with λ_1, since adversaries benefit from the background information. Furthermore, in Fig. 4(d), ϵ decreases with λ_2, sicnce more background information leads to a larger probability that Formula (2) does not hold. Lastly, as analyzed above, Theorem 7 also motivates the designer to select quantified users' data to achieve better performance against inference attacks.

IV. Conclusion

It is argued that k-anonymity suffers from inference attacks. Although there are a number of sophisticated k-anonymity techniques against these attacks, they do not have any rigorous theoretical analysis of the performance of the proposed k-anonymity techniques. In this paper, we have presented the first theoretical foundation that gives a non-asymptotic bound on the performance of k-anonymity against inference attacks. In addition, our work thoroughly and theoretically analyses why k-anonymity is susceptible to these attacks employing the modified Stein method. We believe that it can bridge the gap between design and evaluation, enabling a designer to propose a more practical k-anonymity technique in real-life scenarios to resist inference attacks.

Appendix

A. Proof of Theorem 1

Proof: Since $\mathcal{F}$ selects each tuple in $\mathcal{D}$ with probability p, $\mathcal{F}$ can be mapped to the mechanism $\mathcal{M}$ that meets: $\mathcal{M} = \sum_1^n X_i$, $X_i \sim B(1,p)$, $\mathcal{M} \in [np-\mu, np+\mu]$, $np+\mu < n$, and $np-\mu > k$.

(1) For the statement (I), denote $\parallel \alpha - \alpha^* \parallel_1 = 1$, $\alpha \in [np-\mu, np+\mu]$. We first have to bound $\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]}$. When $0 < p < \frac{1}{2} - \frac{k}{2n}$, $\alpha = \lceil np-\mu \rceil$, and $\alpha^* = \alpha - 1$, $\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]}$ achieves the maximum value. Otherwise, when $\alpha = \lfloor np+\mu \rfloor$, and $\alpha^* = \alpha + 1$, $\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]}$ achieves the maximum value.

Case 1: $0 < p < \frac{1}{2} - \frac{k}{2n}$ and $X_i \sim B(1,p)$. Then we get

$$\begin{aligned}\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]} &= \frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha-1]}\\ &= \frac{n-\lceil np-\mu\rceil+1}{\lceil np-\mu\rceil}\frac{p}{1-p}\\ &< \frac{n-(np-\mu)+1}{np-\mu}\frac{p}{1-p}\end{aligned} \tag{3}$$

Denote $\exp(\epsilon) = \frac{n-(np-\mu)+1}{np-\mu}\frac{p}{1-p}$, and thus we get $\epsilon = \ln(\frac{n-(np-\mu)+1}{np-\mu}\frac{p}{1-p})$.

Case 2: $1 > p > \frac{1}{2} - \frac{k}{2n}$ and $X_i \sim B(1,p)$. Similarly,

$$\begin{aligned}\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]} &= \frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha+1]}\\ &= \frac{\lfloor np+\mu\rfloor+1}{n-\lfloor np+\mu\rfloor}\frac{1-p}{p}\\ &< \frac{np+\mu+1}{n-(np+\mu)}\frac{1-p}{p}\end{aligned} \tag{4}$$

Denote $\exp(\epsilon) = \frac{np+\mu+1}{n-(np+\mu)}\frac{1-p}{p}$. So $\epsilon = \ln(\frac{np+\mu+1}{n-(np+\mu)}\frac{1-p}{p})$.

So far, we have proved for $\alpha \in [np-\mu, np+\mu] \cap \mathbb{Z}$ and $\parallel \alpha - \alpha^* \parallel_1 = 1$, the following holds:

$$Pr[\mathcal{M}=\alpha] < \exp(\epsilon) Pr[\mathcal{M}=\alpha^*], \tag{5}$$

where

$$\epsilon = \begin{cases} \ln(\frac{n-(np-\mu)+1}{np-\mu}\frac{p}{1-p}), & p < \frac{1}{2} - \frac{k}{2n}\\ \ln(\frac{np+\mu+1}{n-(np+\mu)}\frac{1-p}{p}), & p \geq \frac{1}{2} - \frac{k}{2n}\end{cases} \tag{6}$$

Next, we focus on the probability δ that the above inequality is violated. According to the Chernoff bounds, we have

$$Pr[\mathcal{M} < np-\mu] + Pr[\mathcal{M} > np+\mu] \leq 2\exp(-2\mu^2/n) \tag{7}$$

Since $\mathcal{F}$ selects no less than k tuples in $\mathcal{D}$, $\delta = 2\exp(-2\mu^2/n) - \exp(-2(np-(k-1))^2/n)$. We get $\mu = \sqrt{-\frac{1}{2}n\ln[\delta + \exp(\frac{-2(np-k+1)^2}{n})]}$. We further get,

$$\epsilon = \begin{cases} \ln(\frac{n+1}{np - \sqrt{-\frac{1}{2}n\ln[\delta+\exp(\frac{-2(np-k+1)^2}{n})]}} - 1)\\ + \ln\frac{p}{1-p}, p < \frac{1}{2} - \frac{k}{2n};\\ \ln(\frac{n+1}{n-np-\sqrt{-\frac{n}{2}\ln[\delta+\exp(\frac{-2(np-k+1)^2}{n})]}} - 1)\\ + \ln\frac{1-p}{p}, p \geq \frac{1}{2} - \frac{k}{2n};\end{cases} \tag{8}$$

(2) For the statement (II), when $p < \frac{1}{2} - \frac{k}{2n}$, we aim to find the maximum value of α so that $Pr[\mathcal{M}=\alpha] \geq \exp(\epsilon) Pr[\mathcal{M}=\alpha^*]$. Similar to the analysis of statement (I), we have:

$$\begin{aligned}\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]} &= \frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha-1]}\\ &= \frac{n-\alpha+1}{\alpha}\frac{p}{1-p} \geq \exp(\epsilon)\end{aligned} \tag{9}$$

So we get

$$\alpha \leq \frac{p(n+1)}{\exp(\epsilon)(1-p)+p} \tag{10}$$

According to the Chernoff bounds,

$$Pr[\mathcal{M} < np-(np-\alpha)] \leq \exp(\frac{-2(np-\frac{p(n+1)}{\exp(\epsilon)(1-p)+p})^2}{n}) \tag{11}$$

As no less than k tuples are selected, we denote $\delta_1 = \exp(\frac{-2(np-\frac{p(n+1)}{\exp(\epsilon)(1-p)+p})^2}{n}) - \exp(\frac{-2(np-(k-1))^2}{n})$.

When $p \geq \frac{1}{2} - \frac{k}{2n}$, we want to find the minimum value of α so that $Pr[\mathcal{M}=\alpha] < \exp(\epsilon) Pr[\mathcal{M}=\alpha^*]$ does not hold.

$$\begin{aligned}\frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha^*]} &= \frac{Pr[\mathcal{M}=\alpha]}{Pr[\mathcal{M}=\alpha+1]}\\ &= \frac{\alpha+1}{n-\alpha}\frac{1-p}{p}\\ &\leq \exp(\epsilon)\end{aligned} \tag{12}$$

We get $\alpha \geq \frac{\exp(\epsilon)np-(1-p)}{\exp(\epsilon)p+1-p}$.

According to the Chernoff bounds, we get

$$Pr[\mathcal{M} > np+(\alpha-np)] \leq \exp(\frac{-2(\frac{\exp(\epsilon)np-(1-p)}{\exp(\epsilon)p+1-p}-np)^2}{n}) \tag{13}$$

Thus we get $\delta_2 = \exp(\frac{-2(\frac{\exp(\epsilon)np-(1-p)}{\exp(\epsilon)p+1-p}-np)^2}{n})$.

Finally, $\delta = \max\{\delta_1, \delta_2\} = \max\{\exp[\frac{-2(np-\frac{p(n+1)}{\exp(\epsilon)(1-p)+p})^2}{n}] - \exp[\frac{-2(np-(k-1))^2}{n}], \exp[\frac{-2(\frac{\exp(\epsilon)np-1+p}{\exp(\epsilon)p+1-p}-np)^2}{n}]\}$.

Overall, Theorem 1 holds. ■

B. Proof of Lemma 2

Proof: We first prove the bounds of parameters σ and c. The ratio of probabilities that $\mathcal{F}$ outputs the same outcome when the input is $\mathcal{D}$ and $\mathcal{D}^*$ is

$$\mid \ln \frac{\exp(-\frac{1}{2\sigma^2}u^2)}{\exp((-\frac{1}{2\sigma^2})(u+s)^2)} \mid = \mid \frac{1}{2\sigma^2}(2us+s^2) \mid \tag{14}$$

which is bounded by ϵ_1, so we get $u \leq \frac{\sigma^2\epsilon_1}{s} - \frac{s}{2}$. In addition, since more than k data is selected to perform k-anonymity, $\frac{\sigma^2\epsilon_1}{s} - \frac{s}{2} > k$. Moreover, Formula (14) is bounded by ϵ_1 with probability at least $(1-\delta_1)$, and thus

$$Pr[x \geq \frac{\sigma^2\epsilon_1}{s} - \frac{s}{2}] < \frac{\sigma}{\sqrt{2\pi}} \exp(-\frac{(\frac{\sigma^2\epsilon_1}{s}-\frac{s}{2})^2}{2\sigma^2}) < \delta_1 \tag{15}$$

Denote $\sigma = cs/\epsilon_1$. According to Formula (15), we get

$$\ln(c - \frac{\epsilon_1}{2c}) + \frac{1}{2}(c^2 - \epsilon_1 + \frac{\epsilon_1^2}{4c^2}) > \ln \frac{1}{\sqrt{2\pi}\delta_1} \tag{16}$$

Since function $g(\epsilon_1) = \ln(c - \frac{\epsilon_1}{2c}) + \frac{1}{2}(c^2 - \epsilon_1 + \frac{\epsilon_1^2}{4c^2})$ decreases with ϵ_1 and $0 < \epsilon_1 < 1$,

$$\ln \frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln \frac{1}{\sqrt{2\pi}\delta_1} \tag{17}$$

Since $\frac{\sigma^2\epsilon_1}{s} - \frac{s}{2} > k$ and $\sigma = cs/\epsilon_1$, $\frac{c^2s}{\epsilon_1} - \frac{s}{2} > k$.

Then we prove $Pr[u+Q \in \mathcal{O}] \leq \exp(\epsilon_1)Pr[v+Q \in \mathcal{O}] + \delta_1$ as follows.

$$\begin{aligned} Pr[u+Q \in \mathcal{O}] =& Pr[u+Q \in \mathcal{O} \mid u \leq \frac{\sigma^2\epsilon_1}{s} - \frac{s}{2}] \\ &+ Pr[u+Q \in \mathcal{O} \mid u > \frac{\sigma^2\epsilon_1}{s} - \frac{s}{2}] \\ \leq& Pr[u+Q \in \mathcal{O} \mid u \leq \frac{\sigma^2\epsilon_1}{s} - \frac{s}{2}] + \delta_1 \\ \leq& \exp(\epsilon_1)Pr[v+Q \in \mathcal{O} \mid u \leq \frac{\sigma^2\epsilon_1}{s} - \frac{s}{2}] \\ &+ \delta_1 \\ \leq& \exp(\epsilon_1)Pr[v+Q \in \mathcal{O}] + \delta_1 \end{aligned} \tag{18}$$

In summary, Lemma 2 holds. ■

C. Proof of Theorem 4

Proof: Denote $W_i = \mathcal{D}_i - \pi_i$, then $mean(W_i) = 0$, $mean(W_i^2) = \varphi_i^2 - \pi_i^2$, $mean(W_i^3) = \chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3$, and $var(W_i) = \varphi_i^2 - \pi_i^2$. Denote $Q \sim (0, \sum_{i=1}^n(\varphi_i^2 - \pi_i^2))$. According to the Lemma 3, we get,

$$\begin{aligned} Pr[W \in B_u] \leq& Pr[Q \in B_u] \\ &+ \frac{1.1182\sum_{i=1}^n(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}} \end{aligned} \tag{19}$$

According to Lemma 2, we get

$$\begin{aligned} &Pr[Q \in B_u] + \frac{1.1182\sum_{i=1}^n(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}} \\ &\leq \exp(\epsilon)Pr[Q \in B_v] \\ &+ \frac{1.1182\sum_{i=1}^n(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}} + \delta_1 \end{aligned} \tag{20}$$

Let $\delta_1 = \frac{1}{\sqrt{n}}$, from Lemma 2, we get

$$\epsilon_1 \geq \frac{c's}{\sqrt{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)}}, \tag{21}$$

where c' is the minimum value c satisfying:

$$\begin{cases} \ln\frac{2c^2-1}{2c} + \frac{4c^4-4c^2+1}{8c^2} > \ln\frac{\sqrt{n}}{\sqrt{2\pi}}, \\ \frac{c^2s}{\epsilon_1} - \frac{s}{2} > k \end{cases} \tag{22}$$

Thus, we get the privacy parameter ϵ as

$$\epsilon = \frac{c's}{\sqrt{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)}} \tag{23}$$

Again from Lemma 3, we get

$$\begin{aligned} &\exp(\epsilon)Pr[Q \in B_v] + \frac{1.1182\sum_{i=1}^n(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{n}} \\ &\leq \exp(\epsilon)Pr[Q \in B_v] \\ &+ \frac{1.1182\sum_{i=1}^n(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}}(1 + \exp(\sqrt{\frac{s^2\ln n}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)}})) \\ &+ \frac{1}{\sqrt{n}} \end{aligned} \tag{24}$$

So we get privacy parameter δ

$$\begin{aligned} \delta =& \frac{1.1182\sum_{i=1}^n(\chi_i^3 - 3\varphi_i^2\pi_i + 2\pi_i^3)}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)^{\frac{3}{2}}} \times \\ &(1 + \exp(\sqrt{\frac{s^2\ln n}{\sum_{i=1}^n(\varphi_i^2 - \pi_i^2)}})) + \frac{1}{\sqrt{n}} \end{aligned} \tag{25}$$

Lastly, we get

$$Pr[W \in B_u] \leq \exp(\epsilon)Pr[W \in B_v] + \delta \tag{26}$$

So we get

$$Pr[\mathcal{D} \in B_{u'}] \leq \exp(\epsilon)Pr[\mathcal{D} \in B_{v'}] + \delta \tag{27}$$

According to Definition 2, Therorem 4 holds. ■

D. Proof of Theorem 5

Proof: Denote $B_u = \{b + u : b \in B\}$, $B_u/\sigma = \{b/\sigma : b \in B_u\}$, and $Q \sim N(0, n\sigma^2)$. According to Lemma 3,

$$Pr[\frac{S}{\sigma} \in \frac{B_u}{\sigma}] \leq Pr[\frac{Q}{\sigma} \in \frac{B_u}{\sigma}] + 2d_K(\frac{S}{\sigma}, Q), \tag{28}$$

where $d_K(\frac{S}{\sigma}, Q) = sup_{t \in R} \mid F_S(t) - F_Q(t) \mid$ is the Kolmogorov metric [26]. According to Lemma 2,

$$Pr[\frac{Q}{\sigma} \in \frac{B_u}{\sigma}] + 2d_K(\frac{S}{\sigma}, Q) \leq \exp(\epsilon) Pr[\frac{Q}{\sigma} \in \frac{B_v}{\sigma}] + 2d_K(\frac{S}{\sigma}, Q) + \delta_1, \tag{29}$$

According to Lemma 3,

$$\begin{aligned} &\exp(\epsilon) Pr[\frac{Q}{\sigma} \in \frac{B_v}{\sigma}] + 2d_K(\frac{S}{\sigma}, Q) + \delta_1 \leq \\ &\exp(\epsilon) Pr[\frac{S}{\sigma} \in \frac{B_v}{\sigma}] + 2d_K(\frac{S}{\sigma}, Q)(1 + \exp(\epsilon)) + \delta_1 \end{aligned} \tag{30}$$

So, combing Formulas (28) (29) (30), we get

$$\begin{aligned} Pr[\frac{S}{\sigma} \in \frac{B_u}{\sigma}] \leq &\exp(\epsilon) Pr[\frac{S}{\sigma} \in \frac{B_v}{\sigma}] \\ &+ 2d_K(\frac{S}{\sigma}, Q)(1 + \exp(\epsilon)) + \delta_1 \end{aligned} \tag{31}$$

Denote $\delta = 2d_K(\frac{S}{\sigma}, Q)(1 + \exp(\epsilon)) + \delta_1$, we have

$$Pr[\frac{S}{\sigma} \in \frac{B_u}{\sigma}] \leq \exp(\epsilon) Pr[\frac{S}{\sigma} \in \frac{B_v}{\sigma}] + \delta \tag{32}$$

Next, we focus on computing parameters ϵ and δ. δ_1 and ϵ are parameters in Lemma 2. According to the proof of Lemma 2, we let $\delta_1 = \frac{1}{\sqrt{n}}$ and thus get

$$\epsilon = \frac{c' s}{\sqrt{var[\sum_{i=1}^{n} \mathcal{D}_i]}} \tag{33}$$

where c' is the minimum value c satisfying:

$$\begin{cases} \ln \frac{2c^2 - 1}{2c} + \frac{4c^4 - 4c^2 + 1}{8c^2} > \ln \frac{\sqrt{n}}{\sqrt{2\pi}}, \\ \frac{c^2 s}{\epsilon_1} - \frac{s}{2} > k, \end{cases} \tag{34}$$

According to Theorem 3.6 in [26],

$$\begin{aligned} d_K(\frac{S}{\sigma}, Q) &\leq (\frac{2}{\pi})^{1/4} \sqrt{d_W(\frac{S}{\sigma}, Q)} \\ &\leq (\frac{2}{\pi})^{1/4} \sqrt{\frac{\lambda_1^2}{\sigma^3} \sum_{i=1}^{n} \chi_i^3 + \frac{\lambda_1^{\frac{3}{2}} \sqrt{26}}{\sigma^2 \sqrt{\pi}} \sqrt{\sum_{i=1}^{n} \xi_i^4}}, \end{aligned} \tag{35}$$

where d_W is the Wasserstein metric [26]. So we get

$$\begin{aligned} \delta =& 2(\frac{2}{\pi})^{1/4} \sqrt{\frac{\lambda_1^2}{\sigma^3} \sum_{i=1}^{n} \chi_i^3 + \frac{\lambda_1^{\frac{3}{2}} \sqrt{26}}{\sigma^2 \sqrt{\pi}} \sqrt{\sum_{i=1}^{n} \xi_i^4}} \\ &\times (1 + \exp(\epsilon)) + \frac{1}{\sqrt{n}} \end{aligned} \tag{36}$$

In summary, Theorem 5 holds. ■

References

[1] C. Wang, H. Lin, and H. Jiang, “CANS: Towards congestion-adaptive and small stretch emergency navigation with wireless sensor networks,” *IEEE Transactions on Mobile Computing*, vol. 15, no. 5, pp. 1077–1089, 2016.

[2] Q. Yang, D. Li, W. Yu, Y. Liu, D. An, X. Yang, and J. Lin, “Towards data integrity attacks against optimal power flow in smart grid,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1726–1738, 2017.

[3] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, “A survey on security and privacy issues in internet-of-things,” *IEEE Internet of Things Journal*, vol. 4, no. 4, pp. 1250–1258, 2017.

[4] H. Huang, H. Yin, G. Min, H. Jiang, J. Zhang, and Y. Wu, “Data-driven information plane in software-defined networking,” *IEEE Communications Magazine*, vol. 55, no. 6, pp. 218–224, 2017.

[5] H. Huang, H. Yin, G. Min, J. Zhang, Y. Wu, and X. Zhang, “Energy-aware dual-path geographic routing to bypass routing holes in wireless sensor networks,” *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1339–1352, 2018.

[6] G. Min, J. Hu, and M. E. Woodward, “Performance modelling and analysis of the txop scheme in wireless multimedia networks with heterogeneous stations,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4130–4139, 2011.

[7] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[8] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, “Community-enhanced de-anonymization of online social networks,” in *Proc. of ACM CCS*, 2014.

[9] J. Song, S. Lee, and J. Kim, “Inference attack on browsing history of twitter users using public click analytics and twitter metadata,” *IEEE Transactions on Dependable and Secure Computing*, vol. 13, no. 3, pp. 340–354, 2016.

[10] S. Ji, W. Li, M. Srivatsa, and R. Beyah, “Structural data de-anonymization: Theory and practice,” *IEEE/ACM Transactions on Networking*, vol. 24, no. 6, pp. 3523–3536, 2016.

[11] J. Qian, X.-Y. Li, C. Zhang, and L. Chen, “De-anonymizing social networks and inferring private attributes using knowledge graphs,” in *Proc. of IEEE INFOCOM*, 2016.

[12] B. Zhou and J. Pei, “Preserving privacy in social networks against neighborhood attacks,” in *Proc. of IEEE ICDE*, 2008, pp. 506–515.

[13] E. Zheleva and L. Getoor, “Preserving the privacy of sensitive relationships in graph data,” *Lecture Notes in Computer Science*, vol. 4890, pp. 153–172, 2008.

[14] M. Yuan, L. Chen, and P. S. Yu, “Personalized privacy protection in social networks,” *Proc. of the VLDB Endowment*, vol. 4, no. 2, pp. 141–150, 2010.

[15] K. Liu and E. Terzi, “Towards identity anonymization on graphs,” in *Proc. of ACM SIGMOD*, 2008.

[16] J. Cheng, A. W.-c. Fu, and J. Liu, “k-isomorphism: privacy preserving network publication against structural attacks,” in *Proc. of ACM SIGMOD*, 2010, pp. 459–470.

[17] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, “Resisting structural re-identification in anonymized social networks,” *Proc. of VLDB Endowment*, vol. 1, no. 1, pp. 102–114, 2008.

[18] L. Zou, L. Chen, and M. T. Özsu, “k-automorphism: A general framework for privacy preserving network publication,” *Proc. of the VLDB Endowment*, vol. 2, no. 1, pp. 946–957, 2009.

[19] X.-Y. Li, C. Zhang, T. Jung, J. Qian, and L. Chen, “Graph-based privacy-preserving data publication,” in *Proc. of IEEE INFOCOM*, 2016, pp. 1–9.

[20] C. Dwork, A. Roth *et al.*, “The algorithmic foundations of differential privacy,” *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[21] S. Perez, “Twitter partners with ibm to bring social data to the enterprise,” *Available: https://techcrunch.com/2014/10/29/twitter-partners-with-ibm-to-bring-social-data-to-the-enterprise/*, Accessed on Sep. 13, 2017.

[22] “Stanford snap datasets,” *Available: http://snap.stanford.edu/data/index.html*, Accessed on Sep. 13, 2017.

[23] “Cmu datasets,” *Available: http://www.casos.cs.cmu.edu/computational-tools/data2.php*, Accessed on Sep. 13, 2017.

[24] R. Bhaskar, A. Bhowmick, V. Goyal, S. Laxman, and A. Thakurta, “Noiseless database privacy,” in *Proc. of International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 2011.

[25] I. Tyurin, “A refinement of the remainder in the lyapunov theorem,” *Theory of Probability & Its Applications*, vol. 56, no. 4, pp. 693–696, 2012.

[26] N. Ross *et al.*, “Fundamentals of stein’s method,” *Probability Surveys*, vol. 8, pp. 210–293, 2011.