2018 IEEE International Conference on Big Data (Big Data)

DRAKO: Differentially pRivate Algorithm to meet K-anonymity for Online portal service

Kangsoo Jung
dept. Computer Engineering
Sogang University
Seoul, Republic of Korea
azure84@naver.com

Jaewon Kim
dept. Computer Engineering
Sogang University
Seoul, Republic of Korea
jason04132@naver.com

Youngjun Kim
dept. Computer Engineering
Sogang University
Seoul, Republic of Korea
june244@naver.com

Seog Park
dept. Computer Engineering
Sogang University
Seoul, Republic of Korea
spark@sogang.ac.kr

Abstract*-Digital data on the Web are nowadays regarded significant sources of information for marketing and user profiling, etc. However, digital data are risky sources of privacy violation. To address privacy breaches, we can use differential privacy, which has become the de facto standard for privacy protection in statistical databases. However, problems need to be solved, including those related to noise parameter configuration, even before differential privacy can be applied into the real world. In this study, we introduce a linkage attack to identify a user with different nicknames for each subservice on a hue online portal service. In addition, we propose a configuration technique for the upper bound of noise parameter ε to prevent linkage attack. We demonstrate the linkage attack with experiments by using real-world online portal service data. Finally, we validate the proposed configuration technique.

Keywords—differential privacy, linkage attack, anonymization, online portal service

I. INTRODUCTION

Many individuals currently use Web services and voluntarily leave their digital footprints. These personal footprints are collected by companies for use in marketing or user profiling. However, the sharing of personal information may cause unintended privacy violation. To prevent privacy breach, k-anonymity and differential privacy have been proposed but anonymization techniques are limited because they cannot prevent background knowledge attack and differential privacy has difficulty in establishing an appropriate level of noise parameter ε value.

In this study, we present a new user identification attack that can occur in huge online portal sites, such as Naver and Daum, which provide content in various categories, namely, news, movie reviews, blogs, etc. In addition, we propose an appropriate configuration technique for the upper bound of the noise parameter to prevent linkage attacks by applying differential privacy. The proposed technique guarantees the same level of privacy protection as k-anonymization when differential privacy is used.

II. RELATED WORK

Differential privacy is a concept that implies a certain level of privacy protection when noise is inserted into the data. The definition of differential privacy is as follows:

Definition 1. Differential privacy [1]

A randomized function M provides ε-differential privacy if, for all datasets D1 and D2 that differ by one element, all S ⊆ Range(K), i.e.,

$$\Pr[\mathrm{M}(\mathrm{D}_1) \in \mathrm{S}] \leq \exp(\varepsilon) \cdot \Pr[\mathrm{M}(\mathrm{D}_2) \in \mathrm{S}].$$

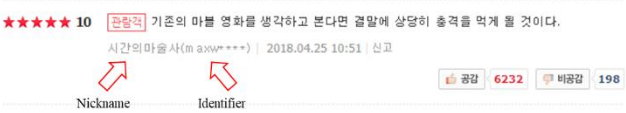
Since the presentation of the concept of differential privacy, there has been an argument over the proper value of the noise parameter ε because no criteria can be used to determine the ε value.

To solve the problem, many studies have been conducted to set the appropriate level of ε [2-6]. The work of [2] showed that the privacy protection level set by an arbitrary ε can be infringed by inferences on previously disclosed information; subsequently, a method for setting ε is proposed on the basis of posterior belief. The factors that need to be considered when setting ε are summarized by [3]. Studies that combine k-anonymization or t-closeness with differential privacy have also been conducted. [4] showed that ε-differential privacy is satisfied if t-closeness is satisfied when $t = max_E \frac{|E|}{N}(1 + \frac{N-|E|-1}{|E|}\exp(\varepsilon))$. Thus, t-closeness and differential privacy can be combined with one another.

III. LINKAGE ATTACK OF HUGE ONLINE PORTAL SERVICE

A. Character of huge online portal service

As previously discussed, huge online portal services, such as Naver [7] or Daum [8], in Korea can set a unique ID for each user upon registration in the portal service. Then, the user can select different nicknames for each subservice, such as for certain reviews (e.g., News or Movie). The user's subservice nickname is presented with the user ID, as shown in Figure 1. However, if the complete ID is shown, then all of the posts written by the user under the different subservices can be linked. To prevent linkage attacks, the portal service provides a privacy policy that prevents accurate ID exposure by masking four "*" characters except the first four characters. Thus, the unique user ID is partially protected, thereby reducing the possibility of overexposing the service user to ID tracking. The ID masking process not only conceals the complete ID but also allows different user IDs to be represented as the same ID. For example, a user ID carmen74 and a user ID carmechanic82 clearly have different IDs.



★★★★★ 10 관람객 기존의 마블 영화를 생각하고 본다면 결말에 상당히 충격을 먹게 될 것이다.

시간의마술사(m axw****) | 2018.04.25 10:51 | 신고

Fig 1: Nickname and identifier of a movie rate service

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE
978-1-5386-5035-6/18/$31.00 ©2018 IEEE

However, when they undergo the masking process, both are represented by the same ID *carm*****. In other words, the masking process reduces the possibility that a specific user is identified because two or more users are provided the same user ID. This approach has the same effect as the privacy protection offered by the existing k-anonymization technique.

B. *Attacker model*

We assume that the administrator has complete knowledge of a user of the portal service, and the full information about the user can be integrated when the administrator generates the required data. All service contents on the portal service are available on the Web. Then, when data are released as mentioned above, an attacker can trace the links between subservice nicknames.

For example, we suppose that there are subservices A and B (Figure 2). An attacker collects user data from service A and service B. The attacker is unable to determine the user who use nicknames Oliver in service A is the same user who use nickname *Blackpanther* or *Kirk* in service B. However, if the portal service administrator provides the total number of each user's post, then the attacker can deduce that *Oliver* and *Blackpanther* may refer to the same user because of the presence of only one combination (e.g., five posts are uploaded by the user). The user does not expect to be identified, and hence, privacy violation is breached.

In this study, we call an attack a "linkage attack" when user identity is deduced by linking nicknames from the different subservices. A linkage attack is defined as follows:

Definition 1. Linkage attack

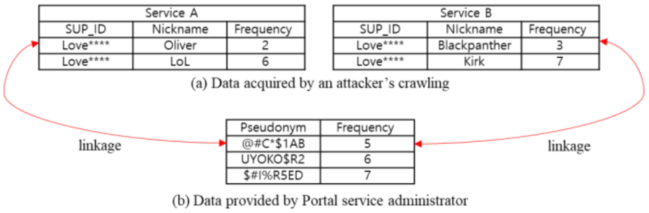
Let user U_i take the nickname sets $UN_i = \{N_{i,1}, ..., N_{i,N}\}$ for subservice Sj ($1 \leq j \leq N$, N = number of subservice). We define the linkage attack by inferring to nicknames $Nick_i$ and $Nick_j$ with the same masking ID for the different subservices that belong to the same user U_k ($Nick_i \in UN_k$, $Nick_j \in UN_k$).

Here, we further propose a configuration technique for the upper bound of the noise parameter.

C. *Noise parameter upper bound configuration with k-anonymization*

C.1 *Necessity of the noise configuration technique.*

An insertion of an appropriate amount of noise is necessary to prevent the linkage attack, and this approach is implemented by applying differential privacy. If the ε value is set too low, then the linkage attack can be prevented but data utility degradation will likely be extensive. On the contrary, if the ε value is set too high, then the linkage attack cannot be prevented. We use the k-anonymity of the ID masking process to set the appropriate level of noise. For example, in Figure 2 (Section 3.2), the total number of cases is six for an attacker who can deduce information by using only his or her collected data.



Service A		
SUP_ID	Nickname	Frequency
Love****	Oliver	2
Love****	LoL	6

Service B		
SUP_ID	NIckname	Frequency
Love****	Blackpanther	3
Love****	Kirk	7

(a) Data acquired by an attacker's crawling

Pseudonym	Frequency
@#C*$1AB	5
UYOKO$R2	6
$#I%R5ED	7

(b) Data provided by Portal service administrator

Fig 2: Example of linkage attack

TABLE I: QUERY RESPONSE PROBABILITY AND POSTERIOR BELIEF

	Query response probability	Posterior belief
Case 1	0.1232	0.8313
Case 2	0.006	0.04
Case 3	0.006	0.04
Case 4	0.001	0.0067
Case 5	0.006	0.04
Case 6	0.006	0.04

In Figure 2, the linkage attack can be regarded successful because the attacker can specify three of the six cases by using public data. The given example also implies that that the attacker should be hindered from specifying any of the six cases to prevent the linkage attack. The implication of the sampled cases is the same as those that use the definition of k-anonymization. Therefore, we propose a noise parameter upper bound configuration technique to satisfy k-anonymity.

C.2 *Noise parameter configuration with k-anonymity.*

When a portal service administrator applies differential privacy to the data, the attacker will infer the original data to the noised data. For example, if we assume that the portal service administrator released noised average number of posts is 9.9 and set the Laplace distribution scale parameter to 1, the probability of 9.9 with its possible combination and posterior belief values are shown in Table I.

The posterior belief equation is

$$\beta(w_i) = \frac{P(Case_i{=}9.9)}{\sum_{j=1}^{6} P(Case_j){=}9.9}. \quad (1)$$

As shown in Table I, although noise is inserted by applying differential privacy, the attacker can infer from the query results that case 1 is highly probable. Thus, when arbitrarily determined, the noise parameter ε cannot prevent the linkage attack described in Section 3.2. We therefore select an appropriate level for noise parameter ε to prevent the linkage attack in the same manner as rendering k-anonymity. The equation for setting the upper bound of noise parameter ε to satisfy k-anonymity [2] is

$$\beta(w_i) = \frac{1}{1+(n-1)e^{\frac{-\varepsilon\Delta v}{\Delta f}}} \leq \rho, \quad (2)$$

where Δf is sensitivity, ρ is the maximum threshold of probability in which a specific case will be inferred (i.e., 1/k based on k-anonymity), n is the total number of possible combinations, and Δv is the maximum value of the difference between the query results for the possible cases.

$$\Delta \text{v} = max_{1\leq i,j\leq n}|f(case_i) - f(case_j)| \quad (3)$$

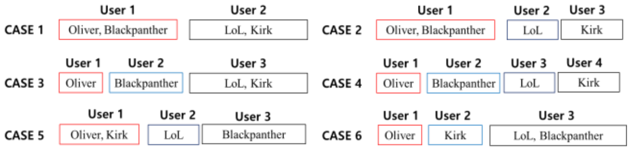


Fig 3: Example of number of cases

TABLE II NUMBER OF USERS WITH THE SAME MASKING ID

	k = 1	k = 2	k = 3	k = 4	4<k
News	156,623	99,626	78,036	61,776	1,538,871
Movie	112,585	76,712	54,075	42,136	859,646

TABLE III NUMBER OF MASKING ID ACCORDING TO NUMBER OF POSSIBLE COMBINATIONS

	n = 3	n = 4	n = 5	n = 6	6<n
News and Movie	13,080	7,598	4,041	9,987	37,776

TABLE IV NUMBER OF THRESHOLD VIOLATIONS AND MASKING ID GROUPS

Original	Proposed ε	ε = 0.1	ε = 2
11,349	1,467	3	3,346

TABLE V RMSE ACCORDING TO E VALUE

	Proposed ε	ε = 0.1	ε = 2
RMSE	175.58	990.33	51.75

Equation (2) is rearranged with respect to ε as follows:

$$\varepsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)\rho}{1-\rho} \quad (4)$$

With equation (4), we can set the upper bound of the noise parameter ε and prevent linkage attack with 1/k probability.

IV. EXPERIMENTAL RESULTS

We collected 49,809,574 comments on News service and 3,164,724 user ratings on Movie review service in Naver, an online portal site of South Korea. The coverage of the experiment is from May 2016 to April 2017. The total number of users for the News service is 1,935,332 while that of for the Movie review service is 1,145,154.

Table II presents the distribution of users recognized as having the same IDs by the masking process. Table III shows the number of masking ID according to the number of possible combinations with commonly masked IDs for the News service and the Movie review service. *Comparison of arbitrary ε and proposed ε for accuracy and privacy protection*

A comparative evaluation is conducted for the proposed ε and the arbitrary ε in terms of accuracy and privacy protection. We perform an averaging of the number of post queries and repeat the calculation ten times. Then, the proposed ε is compared with ε = 0.1 and ε = 2. In this experiment, we set the posterior belief threshold to less than 0.5. Thus, the privacy protection level can be evaluated as the number of masking ID groups with posterior probability greater than 1/2.

A total of 11,349 masking ID groups violate the threshold of the original data whereas 3,346 masking ID groups violate the threshold when ε = 2 (Table IV). When ε = 0.1, there will be no violation, but data utility for privacy protection is significantly reduced. In the proposed ε value, only 1,467 masking ID groups violate the threshold; thus, it is better than the original or ε = 2 cases.

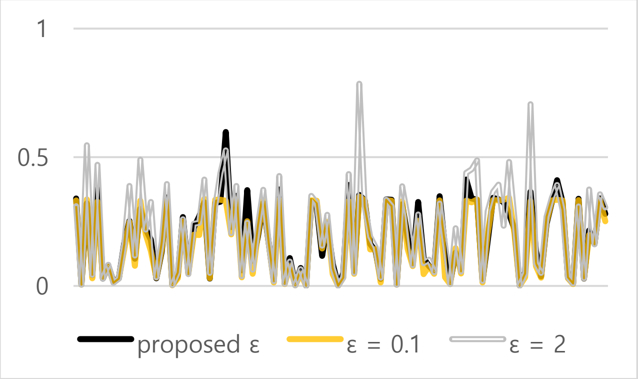


Fig 4: Posterior belief graph according to ε value

Table V shows the RMSE values for the proposed ε, ε = 0.1, and ε = 2. Previously, in Table IV, ε = 0.1 provided the perfect privacy protection, but it will significantly deteriorate data utility. The RMSE value is less than 1 when ε = 2, but this case implies a breach in privacy protection.

The experimental result shows that the configuration technique proposed by the present work succeeds in setting the appropriate ε value and thus can realize both accuracy and privacy protection.

V. CONCLUSION AND FUTURE WORK

In this study, we proposed a novel linkage attack problem in which a user with different nicknames for each subservice can be identified by linking. We presented a configuration technique for the upper bound of the noise parameter to provide k-anonymity levels of identification protection. We verified that the proposed configuration technique can be used to set the appropriate upper bound of the noise parameter.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)(2017-0-00498, A Development of Deidentification Technique based on Differential privacy)

REFERENCES

[1] C. Dwork, Differential privacy: A survey of results, Proceedings of the International Conference on Theory and Applications of Models of Computation, pp.1-19, 2008.

[2] J. Lee, C. Clifton, How Much is Enough? Choosing Epsilon for Differential Privacy, Proceedings of the International Conference on Information Security, pp. 325–340, 2011.

[3] J. Hsu, et al, Differential Privacy: An Economic Method for Choosing Epsilon, Proceedings of the 27th IEEE Computer Security Foundations Symposium, pp.1–29, 2014.

[4] J. Soria-Comas, J. Domingo-Ferrert, Differential privacy via t-closeness in data publishing, In Privacy, Proceedings of the International Conference on Security and Trust, pp. 27-35, 2013.

[5] N. Holohan, et al. (k, ε)-Anonymity: k-Anonymity with ε-Differential Privacy. arXiv preprint arXiv:1710.01615, 2017.

[6] N. Li, et al. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, pp. 32-33, 2012.

[7] Never, www.naver.com

[8] Daum, www.daum.net