(δ,l)-diversity: Privacy Preservation for Publication Numerical Sensitive Data

Mohammad-Reza Zare-Mirakabad

Department of Computer Engineering
Scool of Electrical and Computer
Yazd University, Iran
mzare@yazduni.ac.ir

Abstract.

(ε,m)-anonymity considers ε as the interval to define similarity between two values, and m as the level of privacy protection. For example {40,60} satisfies (ε,m)-anonymity but {40,50,60} doesn't, for ε=15 and m=2. We show that protection in {40,50,60} sensitive values of an equivalence class is not less (if don't say more) than {40,60}. Therefore, although (ε,m)-anonymity has well studied publication of numerical sensitive values, it fails to address proximity in the right way. Accordingly, we introduce a revised principle which solve this problem by introducing (δ,l)-diversity principle. Surprisingly, in contrast with (ε,m)-anonymity, the proposed principle respects monotonicity property which makes it adoptable to be exploited in other anonymity principles.

Keywords:

k-anonymity, privacy preservation, (ε,m)-anonymity, monotonicity, proximity

1. Introduction

Privacy protection of personal data has become a serious concern in recent years. Organizations want/need to publish operational data for the purpose of business visibility and effective presence on the World Wide Web. Individuals also publish personal data in the hope of becoming socially visible and attractive in the new electronic communication forums. While this data sharing has many benefits, privacy of individuals may be compromised. Specifically data holders are worry about protection against privacy attacks by re-identification, cross referencing and joining on other existent data. Then protecting privacy of individuals has become an important concern by organizations and governments.

Among various approaches addressing this issue, k-anonymity and l-diversity models have recently been studied with considerable attention. k-anonymity [1,2] has been proposed to protect identification of individuals in the published data. Specifically in k-anonymity, data privacy is protected by ensuring that any record in the released data is indistinguishable from at least (k-1) other records with respect to the quasi-identifier, i.e. sets of attributes that can be cross-referenced in other sources to identify objects. Each equivalence class of tuples (the set of tuples with the

Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 535–543, 2012.
 DOI : 10.5121/csit.2012.2352

same value for the attributes in the quasi identifier) has at least k tuples. An individual is hidden in a crowd of size k, thus the name k-anonymity. Subsequent works on k-anonymity mostly propose algorithms for k-anonymization [3,4].

While k-anonymity prevents identification, l-diversity [5] aims at protecting sensitive information. This is achieved by ensuring that sensitive attribute values are "well represented"' as per the l-diversity principle enounced in [5]. Actually this principle is stronger than k-anonymity since can protect private information from being disclosed.

Although almost all of the l-diversity principles consider both categorical and numerical sensitive information, they fail to adequately protect numerical sensitive attributes. More exactly the information breach can be occurred if an adversary could infer that sensitive value of an individual is in a short interval with high confidence.

Consider Figure 1 as an example that shows a generalized data and salary of a fictitious company. This table satisfies diversity based on all diversity principles point of view. Especially it is distinct 3-divers and even frequency 3-diverse. It is because every equivalence class with respect to (Age,Zip) has at least 3 distinct values and none of them is repeated more than one in each equivalence class(for the case of frequency l-diversity).

Age	Zip	Salary (K)
[17,25]	**11*** **	**490**
[17,25]	**11*** **	**500**
[17,25]	**11*** **	**510**
[17,25]	**11*** **	**1000**
[26,35]	***11****	***500***
[26,35]	***11****	***600***
[26,35]	***11****	***700***
[36-45]	**11*** **	**1000**
[36-45]	**11*** **	**510**
[36-45]	**11*** **	**680**

Fig. 1. 3-diverse employees' data

Enforced by frequency l-diversity (as an example), if an adversary knows an individual's Age and Zip that exists in this table, she can not infer her exact salary with probability more than 1/3. For instance if Alice is 19 years old living in Zip=11700 area and exists in this table, then an adversary only knows she is in first equivalence class. Therefore with probability more than 1/3 her exact salary can not be revealed. However, an attacker can conclude with probability 1 (absolutely confidence) that Alice's salary is very close ("similar"') to 500K (precisely speaking in the range [490K,510K]) which is sufficient for him to reveal her salary.

This problem has recently been addressed by Li et al. [6] as proximity privacy for numerical sensitive data. They propose a new principle, named (ε,m)-anonymity, to eliminate proximity breach for publishing numerical sensitive attributes. Actually if two numerical values are "similar"' (considering an interval expressed by parameter ε) they are assumed as identical value in the term of diversity. Hence, it provides more robust protection to enforce diversity of sensitive values in each equivalence class. Precisely, they consider an interval neighborhood for numerical values as follows:

Consider Table T containing tuples t with sensitive attribute S. Absolute and relative ε-neighborhood interval for each tuple t are defined as [t.S-ε, t.S+ε] and [t.S (1-ε), t.S (1+ε] respectively where ε is any non-negative value in former and a real value in range [0,1] in the latter. In terms of similarity, they consider two interpretations. The first one expresses that two values x and y are similar if their absolute difference is at most ε, i.e. $|y-x| \leq \varepsilon$. Another consideration is similarity in a relative sense. That is y is similar to x, if $|y-x| \leq \varepsilon.x$. These two interpretation of similarity in (ε,m)-anonymity result to absolute and relative (ε,m)-anonymity respectively.

The risk of proximity breach of t in each equivalence class E with respect to its quasi identifier is x/|E|, where x is number of tuples in E whose sensitive value falls in ε-neighborhood interval of t. Although their principle can protect against proximity privacy by considering ε-neighborhood and "similarity"', it, however, can not address the similarity in right way. More exactly, what it shows about privacy breach in some equivalence classes is different from what one expect and believe about it. For example base on their definition if one knows sensitive value of an individual is in {40,60} is more anonymous than it is in {40,50,60}. Intuitively it is meaningless. Also their proposed principle lacks monotonicity property which is a prerequisite for exploiting efficient pruning for computing generalization in almost all anonymization algorithms.

In this paper we propose another model, (δ,l)-diversity, which is tackling both these drawbacks. It is exactly conformable with what one imagines about proximity on numerical sensitive data. It also has monotonicity property that can be used to introduce efficient algorithms by exploiting pruning paradigm during generalization process.

The remainder of this paper is organized as follows. In section 2 we survey related work with a focus on l-diversity and necessity of special attention on numerical sensitive data. In section 3 we address details of the problem and the defects of previously proposed principle. We bring definitions of necessary notions and our proposed principle in section 4. Section 5 is dedicated to the algorithm for checking (δ,l)-diversity condition. Finally we conclude in section 6 with directions to future works.

2. Literature Review

l-diversity [5] aims at protecting sensitive information. It guarantees that one cannot associate, beyond a certain probability, an object with sensitive information. This is achieved by ensuring that values of sensitive attributes are "well represented" as per the l-diversity principle enounced in [5].

Iyengar [7] characterizes k-anonymity and l-diversity as identity disclosure and attribute disclosure, respectively. Actually this principle is stronger than k-anonymity [1,2] since can protect private information from being disclosed. Many different instances of this principle, together with corresponding transformation processes, have been proposed. For instance distinct l-diversity [8], entropy l-diversity and recursive (c,l)-diversity [5], (α,k)-anonymity [9], and t-closeness [8] are some of the proposed instances (usually presented with the corresponding diversification algorithms).

The authors of [10] present an instance of l-diversity, as a trade-off between other instantiations, such that in each equivalence class at most a 1/l fraction of tuples can have same value for the sensitive attribute. This definition is most popular in recent works like [11]. We refer to this as "**frequency l-diversity**". (α,k)-anonymity, introduced in [9] uses similar frequency requirements to selected values of the sensitive attributes known to be sensitive values.

Confusingly, the name l-diversity is sometimes used by authors to refer to any of the above instances rather than to the general principle.

Recently authors of [6] have considered risk of proximity breach in publishing numerical sensitive data. They survey most of known anonymization principles and show inadequacy of them in preventing proximity breach, even if an expected level of anonymity has been enforced.
Anonymity principles can be divided to two groups, according to whether they are designed for categorical sensitive attributes or numeric ones. One group of principles addressing categorical sensitive attributes such as l-diversity [5] and its variants, (c,k)-safety [12], and Skyline-privacy [13] are shown have common weakness with respect to proximity privacy. This is because they consider "different values", no matter they are close to each other or not, which have not any sense of proximity. This consideration is somewhat reasonable for categorical sensitive values. It is not, however, appropriate for numerical values which are different by a very small difference.
Also another group, although addressing numerical sensitive attributes, has some limitation for preventing proximity breaching as well. They show principles like (k,e)-anonymity [14] suffer from proximity breaching. Even Variance Control and t-closeness [8], which target numerical sensitive values and try to retain distribution of sensitive attribute of overall table in every equivalence class, can not completely solve the problem. δ-presence [15] is only one option for protecting proximity attacks but only for the case that attacker is not sure about the existence of the victim individual in the data. This assumption is not realistic in many applications which an individual definitely exists in the dataset and an adversary only try to reveal the sensitive information.

Regarding inadequacy of all these previous anonymization principles, [6] introduces a new principle, (ε)-anonymity to eliminate proximity breach in publishing numerical sensitive values.

3. Problem Statements

Example 1. Consider two equivalence classes E1 and E2 containing sensitive values {40,60} and {50,80} respectively.

3.1 Inadequacy of (ε,m)-anonymity

Consider Example 1, especially equivalence class E1 containing two tuples with sensitive values {40,60}. According to (ε,m)-anonymity, E1 fulfill (ε=15,m=2)-anonymity property. As ε=15, one can conclude probability of "t.S is similar to 40" is 1/2 because, as we already explained, its ε-neighborhood interval contains only one value (40 itself). The same result is for 60. However the probability of "t.S is similar to 50" is 1. Because ε-neighborhood interval for 50 contains two values (40 and 60). It shows although (ε,m)-anonymity, with ε=15 and m=2, is met for values included in equivalence class, it may fails to protect against some useful inferences by attacker. In

this example the proximity breach is occurred for this equivalence class with 100% confidence, for the inference "the value is in [40,60]", although the probability is 1/2 for "value is 40 or 60".

To show the weakness of (ε,m)-anonymity, assume also 50 exists in sensitive values of equivalence class E1, i.e. E1 includes sensitive values {40,50,60}. Now for ε=15, m is bounded to 1, because ε-neighborhood interval for 50 contains all these three values, hence m=3/3 equal to 1. From the privacy presentation point of view, smaller the m, less privacy preservation.
In sum, it is intuitive that if sensitive value of an individual lies in a group including {40,50,60} is more protected than individual with sensitive value in {40,60}. Understanding that sensitive value, say Salary, of an individual is 40K or 60K is not only safer than understanding it is 40K, 50K or 60K, but also the latter one is more confusing and anonymous. The (ε,m)-anonymity, however give higher level an anonymity for the former. This shows an intuitive and implicit drawback exists in this model.

It needs a different property to take this kind of inference into account and overcome this drawback.

3.2 Lack of monotonicity property

Against all other privacy preservation principles, (ε,m)-anonymity has not monotonicity property. Actually this property says "if two equivalence classes E1 and E2 satisfy a principle condition, their union (E1∪E2) also satisfies this principle". Most of anonymization principles exploit this property in generalization process to check stopping condition and prune search tree to prevent extra generalization. [6] shows this property is not supported by (ε,m)-anonymity. In can be shown by a simple counter-example as follows [6]:

Consider Example 1 again. For ε=15 and m=2, both of them fulfill (ε,m)-anonymity. However their union {40, 50, 60, 80} doesn't satisfy (ε=15,m=2)-anonymity. It is because for tuple with t.S=50, ε-neighborhood interval is [35, 65] including 3 values (40, 50 and 60). Then the risk of proximity breach is 3/4 that is more than 1/2 (1/m). Then for ε=15, m=2 the property is violated for union equivalence class.

The lack of monotonicity property not only prevents exploiting pruning paradigms during generalization process but also restricts this principle to be adopted and employed by other principles.

3.3 Contribution

In sum, the definition, notion and solution proposed in [6] suffers from 2 drawbacks. One drawback is that it can not show the exact insight and practical protection which is supposed to express by definition.

Another drawback comes from the lack of monotonicity property which is the prerequisite of an efficient top-down pruning algorithm for computing generalization.

Motivated by these drawbacks of (ε,m)-anonymity, we are proposing another model, named (δ,l)-diversity, in the manner to overcome both drawbacks. It is completely consistent and more regular base on what data holder expect and suppose about proximity privacy. It simultaneously

possesses monotonicity property. By introducing such a property, not only we support a new aspect of privacy preservation for publishing numerical sensitive values, but also it can be adopted to be employed in previous anonymization principles.

4. Definitions

l-diversity is defined with respect to sensitive attributes. Without loss of generality we consider a single sensitive attribute. In this paper we write r(Q, s) to refer to the instance r of R in which $s \in R$ is the sensitive attribute, $Q \subseteq R$ is the set of non-sensitive attributes and $s \notin Q$. Frequency l-diversity requires that each value of the sensitive attributes in each equivalence class E (sets of tuples that "have the same values for the attributes in Q") appear at most |E|/l times in E.

Definition 1 (Frequency l-diversity [10]). Frequency l-diversity is enforced by a given equivalence class E, if for every sensitive value v in E at most 1/l of the tuples in E have sensitive value "equal" to v.

Definition 2 ((ε,m)-anonymity [6]). (ε,m)-anonymity is satisfied by a given quasi identifier group G, if for every sensitive value x in G at most 1/m of the tuples in G have sensitive value "similar" to x. (x and y are similar if $|y-x| \leq \varepsilon$.)

A consequent result of this definition is: No similar sensitive value appears more than |G|/m times in G and it means:

$$n(x) \leq \frac{|G|}{m} \Rightarrow \mathrm{m} \leq \frac{|\mathrm{G}|}{\mathrm{n(x)}} \qquad (1)$$

where n(x) is number of tuples in G having sensitive value similar to x.

To find m satisfied with a given quasi-identifier group G, one has to find minimum m. Minimum m is occurred by maximum value of n(x). Then we have

$$m = \frac{|G|}{\begin{array}{l}\textit{ma}\text{ximum number of tuples in G}\\ \text{having similar sensitive value}\end{array}}$$

(note that value of m for entire dataset is the minimum value between m values of groups)

Example 2. Consider table in Figure 2 with two generalized groups and numeric sensitive value S. Moreover assume ε=15.

QId	S
G1	40
	60
G2	40
	50
	60

Fig. 2. An example table

For G1, I(40)={40}, then n(40)=1. I(60)={60}, then n(60)=1. Hence m=2/1=2 ($|G|/n(x)_{max}$).

For G2, I(40)={40,50}, then n(40)=2. I(50)={40,50,60}, then n(50)=3. I(60)={50,60}, then n(60)=2. Hence m=3/3=1.

We use notation δ and l instead of ε and m. Also we use term diversity instead of anonymity since intuitively this principle is one variety of l-diversity. Then we name our proposed principle **(δ,l)-diversity**. We use the similar terminology but instead of considering only sensitive values in each equivalence class we consider all values in the δ-interval of them. Then similarity of two values is defined base on **overlapping** of these intervals.

Definition 3 (δ-interval). For each sensitive value v the δ-interval of v is [v-δ,v+δ].

Definition 4 (δ-similarity). Two sensitive values v1 and v2 are δ-similar if their δ-interval is overlapping.

Definition 5 ((δ,l)-diversity). (δ,l)-diversity is satisfied by a given equivalence class E, if for every sensitive value v in E at most 1/l of the tuples in E have sensitive value δ-similar to v.

If we compare this definition with frequency l-diversity, they are exactly same, only implying "δ-similarity" for comparing values, instead of using "equality" in frequency l-diversity.

A consequent result of this definition is: No δ-similar sensitive values appear more than |E|/l times in E and it means:

$$n(v) \le \frac{|E|}{l} \Rightarrow l \le \frac{|E|}{n(v)} \qquad (2)$$

where n(v) is number of tuples in E having sensitive value δ-similar to v.

To find m satisfied with a given equivalence class E, one has to find minimum l. Minimum l is occurred by maximum value of n(v). Then we hav

$$l = \frac{|E|}{\begin{array}{l}\text{maximum number of tuples in E}\\ \text{having } \delta\text{-similar sensitive value}\end{array}}$$

(note that value of l for entire dataset is the minimum value between l values of equivalence classes.)

Example 3. Again consider table in Figure 2 and assume δ=15. In G1, by our definition, 40 is δ-similar to 60 because δ-interval for 40 and 60 are [25,55] and [45,75] respectively. These two intervals overlap, then their respective values are δ-similar. Therefore l for this equivalence class is 2/2=1 and is different from m of (ε,m)-anonymity.

Our definition has two benefits. Firstly it overcomes drawback of (ε,m)-anonymity which can not show exact proximity breach in some cases. Secondly surprisingly this definition honors monotonicity property which all other anonymization principles satisfy as well. The result of this

property is that one can exploit previous proposed generalization algorithms to find (δ,l)-diversity of data while is not possible for (ε,m)-anonymity

5. Checking (δ,l)-diversity

To check whether given dataset is satisfying demand level of anonymity, which is enforced by (δ,l)-diversity, each equivalence classes need to satisfy this property. Assume t is the dataset list and tuples in each equivalence class E have been sorted in ascending order of their sensitive values. We give the algorithm for checking in Figure 3. The checking is carried out in O(|E|).

6. Conclusions

(ε,m)-anonymity considers ε as the interval to define similarity between two values, and m as the level of privacy protection. We showed two drawbacks of this principle including a) it can not show the proximity rightly and b) it lacks of monotonicity property. We revised the definition and proposed another principle, called (δ,l)-diversity which 1) solves the problem exist in ε as similarity interval; 2) in a manner that respects monotonicity property to be adoptable for other principles.

```
Algorithm d-l-checking(E, d, l)
   i=0; j=1; x=0; lE = ∞
   while(j < |E|)
      while(t_i+d < t_x-d)
         i++;
      while(j < |E| and t_j-d ≤ t_i+d)
         j++;
      lNext = |E| /(j-i);
      if (lNext < lE)
         lE = lNext;
      x++;
      if lE ≥ l return True
       return False
```

Fig. 3. Checking (δ,l)-diversity property

We are now working on the anonymization methods (more exactly l-diversification one) to introduce an algorithms for (δ,l)-diversity principle. Actually this algorithm is not as simple as other l-diversity principles, such as frequency l-diversity. It needs more consideration because finding the best equivalence classes, with less information loss and meantime more data utility, base on proposed principle (intervals overlapping as the similarity notion) is not so straightforward.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy, "International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, Tech. Rep., 1998.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering," in Principles of Database Systems(PODS), Chicago, Illinois, USA, 2006.

[4] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," in 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in IEEE 22nd International Conference on Data Engineering (ICDE'06), 2006.

[6] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in ACM Conference on Management of Data (SIGMOD), Vancouver, BC, Canada, 2008, pp. 473–486.

[7] V. Iyengar, "Transforming data to satisfy privacy constraints," in SIGKDD, 2002, p. 279288.

[8] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in IEEE 23rd International Conference on Data Engineering (ICDE), Istanbul, 2007, pp. 106–115.

[9] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(alpha,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2006.

[10] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in Very Large Data Bases (VLDB) Conference, Seoul, Korea, 2006, pp. 139–150.

[11] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in Very Large Data Bases (VLDB) Conference. Vienna, Austria: ACM, 2007.

[12] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in International Conference on Data Engineering (ICDE), 2007.

[13] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy skyline: Privacy with multidimensional adversarial knowledge," in VLDB 07. Vienna, Austria: ACM, 2007.

[14] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," in International Conference on Data Engineering (ICDE), 2007, pp. 116–125.

[15] M. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in ACM SIGMOD International Conference on Management of Data, Beijing, China, 2007.