Read the typeinst.doc document from Springer!!!

Page setup, margins: 5.2 X 4.4

Review: l-diversity – privacy beyond k-anonymity

Comment: Style: title

Panos Vassiliadis

University of Ioannina, Department of Computer Science,
45110 Ioannina, Hellas
pvassil@cs.uoi.gr

Comment: Observe the styles: author, address etc

Abstract. This survey intends to summarize the paper [MaGK06] with a critical point of view. The paper deals with possibilities of attacking the k-anonymity generalization method and provides a method to circumvent potential problems. In this survey, we present the fundamental ideas of the paper and its main contributions, as well as a critical view with respect to the strong and weak aspects of the proposed method.

Comment: Style:abstract

Keywords: privacy, k-anonymity, l-diversity

1 Introduction

Comment: Style: **h**eading1. ATTN: different from **H**eading 1

Privacy in the field of data management deals with the problem of concealing sensitive information about individual records. The main technique explored by the research literature follows the method of domain generalization. The ultimate goal is to conceal each individual tuple into an appropriately constructed group of data, in a way that an attacker cannot easily reason about the participation of individuals into the group.

Comment: Style: p1a (always after a heading, no special indentation of the paragraph's 1st line

Take for example the case of medical records of a relation *T(Name,Age,ZipCode,Disease)* that is to be exported to analysts for data mining purposes. On the one hand, our aim is to provide the analysts with as much statistically important information as possible; on the other hand, we want to hide the relationship of individuals (identified by the *identifier* attribute *Name*) with the *sensitive* attribute *Disease*. This equilibrium among goals is primarily achieved by removing the statistically insignificant attribute *Name* from the published version of the relation. Unfortunately, it is still possible to breach the individuals' privacy via *quasi-identifier* attributes (in our example, *Age* and *ZipCode*) which can convey contextual information to an attacker about the concealed identifier attributes and their linkage to sensitive attributes (in our example, a patient's neighbor who knows the zip code and age of a patient can reason on the patient's disease if there are no other patients with similar characteristics). To this end, these attributes are partially anonymized (e.g., *ZipCode=45110* is generalized to *ZipCode=451***), in order to form larger groups of tuples with the same quasi-identifier values.

Comment: Style: Normal. **Observe** how "variables" are consistently formatted in italics throughout all the text

Previous work ([Sama01] and especially [LeDR05]) has provided an efficient method to compute a value *k* such that each individual tuple falls inside an anonymized group of at least *k* tuples with the same quasi-identifier values. The paper

Comment: Observe how we **cite related work**!

under review [MaGK06] builds upon the previous results by identifying weakness and proposing remedies for the case of k-anonymity. The essence of the contribution of Machanavajjhala, Gehrke, and Kifer in [MaGK06] concerns:

(a) the identification of how the hiding of individual tuples in groups of size k can be breached if the statistical behavior of the groups is biased towards a single, or a few values that are easy to be compromised, and,

(b) the proposal of a new anonymization criterion, l-diversity: Given an anonymized data set comprising a set of quasi-identifier and sensitive attributes, the authors propose that a group must possess at least l "well represented" values for the sensitive attribute in order to safely guarantee privacy from background knowledge of the attackers

The authors have assessed their method by conducting experiments concerning its efficiency and effectiveness over the adult and the lands end data sets.

Comment: Intentionally avoided a paragraph with the roadmap (structure) of the document. Observe the structure anyway…

2. Core of the approach and assessment

Comment: heading2

2.1 Essential l-diversity

The main idea of the paper is to go beyond k-anonymity in ensuring that identifier attributes are not linked to their sensitive counterparts via background knowledge of the attacker. The two highlighted vulnerabilities of k-anonymity are (a) the possibility of a whole group to have the same sensitive value and (b) the possibility of having too few sensitive values in the same group. In both cases, the individuals are not 'hidden in the crowd' of their group since all (or, a large number of) the members of the group have the same sensitive value. If this is the case, if an attacker relates an individual with a certain group, then he can confer with high probability the sensitive values of the hidden individual.

L-diversity is a criterion that tells us whether a group is versatile enough in order to effectively hide its members by exploiting both a large number of members and a large number of 'well-represented' values. The purpose is that the probability of relating an individual with its sensitive values is low, even in the case where the attacker can identify the individual's group. The authors of [MaGK06] propose three ways to implement the term 'well-represented':

(i) the distinct number of sensitive values in a group should be higher than *l*

(ii) the entropy of each group should be higher than *log*(*l*)

(iii) recursive l-diversity is achieved for each group. Assume that we sort the values of an sensitive attribute by their frequency in the group; let r_1, r_2, …, r_m be the respective frequencies. In this case, we require that the highest frequency (r_1) is not greater than the sum of the lowest [*l*..*m*] frequencies (r_l, …, r_m), multiplied by a scale factor *c*. (In other words, the frequent values are not too frequent and the infrequent values are not too infrequent).

Observe: page numbers!

2.2 Experiments, experimental method and findings

The authors have experimented with the well-known adult data set as well as with an otherwise unidentified lands end database. The important characteristics that the authors measure are (a) efficiency and (b) effectiveness (also referred to as 'utility' in the paper).

Efficiency is measured as the time needed to anonymize the data set, by varying the parameters of the anonymization. The opponent method was k-anonymity and the variables *k* and *l* varied from 3 to 7. The performance of the two methods was very similar.

Effectiveness is measured via three metrics (a) generalization height (i.e., number of generalization steps performed), (b) average group size and (c) discernability, defined as the number of non-distinguishable tuples. Again, the parameters that were varied were *k* and *l* and k-anonymity was compared against entropy-based diversity and recursive diversity. The results for the three opponents are surprisingly very similar, with k-anonymity being the winner in several cases (esp., as the group size/value diversity grow).

3. Critical view to the paper

3.1 Strong points

The paper has been highly influential and a point of reference for subsequent works due to its fundamental observations.

1. The first, important observation which highlighted that k-anonymity has flaws has practically opened a whole new ground of research.
2. The observation of diversity inside the groups is also very important: the paper reveals that information hiding is not a matter of hiding an individual tuple in a voluminous group, but also, of guaranteeing that the group is diverse enough to *minimize the probability* to infer the hidden identity.
3. For the largest of its part, the paper is well-written and explanatory.

3.2 Weak points

In the following, we list a couple of points that we find weak in the paper:

1. In terms of technical weaknesses, subsequent research [LiLV07] has proved that privacy via *l*-diversity can be compromised in the case where the sensitive value distributions in a group significantly diverge from the distribution of the values in the whole table.
2. The relationship of the proposed method to the tradeoff of anonymization and suppression is quite unclear. In the experimental section, the authors mention that they did not perform any tuple suppression; however, it is

unclear what happens if suppressions are allowed. Also, the effect of column suppression is not obvious (although it should be noted that the suppressed columns in the experiments are the ones with the smallest domains, which suggests that the results would probably not change much if these columns were present in the computations).

3. Computing l-diversity is obviously hard – even for its simplest form. This is apparent by looking at the experiments, where the performance of the anonymization process is not very satisfactory (can rise up to 25 minutes for a 4M rows data set). The authors do not present any algorithmic results on the efficient computation of the anonymized data set.
4. In terms of presentation, the intuition for entropy and recursive l-diversity as well as Bayes' optimal privacy is not well explained. Neither the intuition, nor the mechanics of these criteria are presented in a way which is very clear to the reader. Also, in the experimental section, there is no discussion of important parameters and data set characteristics in the paper.

3.3 Possibilities for improvement

Clearly, the aforementioned weaknesses can be exploited for future research. With the benefit of retrospection [LiLV07], we know that there are more threats to security than *l*-diversity can block. The efficient computation of *l*-diverse groups is not obvious. The combination of *k*-anonymity and *l*-diversity is also unclear with respect to its safety guarantees.

References

Comment: Pay attention on the citations & read the manual I have given you!
The only deviation from the LNCS style is that instead of numbers, we use the 4+2 notation. Stick with it. Again: **RTFM!!**

[LeDR05] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2005), pages 49–60, Baltimore, Maryland, USA, June 14-16, 2005.

[LiLV07] N. Li, T. Li, S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007), April 15-20, 2007, The Marmara Hotel, Istanbul, Turkey

[MaGK06] A. Machanavajjhala, J. Gehrke, and D. Kifer. l-diversity: Privacy beyond k-anonymity. Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006), 3-8 April 2006, Atlanta, GA, USA.

[Sama01] P. Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE), 13(6), pages 1010–1027, 2001.