Review: l-diversity – privacy beyond k-anonymity

Summary. This report summarizes [MaGK06], a paper that deals with possibilities of attacking the k-anonymity generalization method and provides a method to circumvent potential problems. 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.

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 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) 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) 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 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.

Strong points
The paper has been highly influential and much cited by subsequent works due to its fundamental observations.
1. The first, important observation which highlighted that k-anonymity has flaws has practically opened a new whole 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.

Weak points
In the following, we list a couple of points that we find weak in the paper:
1. In terms of presentation, the intuition for entropy and recursive l-diversity is not very explained. Neither the intuition, nor the mechanics of these criteria are presented in a way which is very clear to the user.
2. 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.
3. 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.

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

