Efficient Recommendation of De-identification Policies using MapReduce

ABSTRACT:

Many data owners are required to release the data in a variety of real world application, since it is of vital importance to discovery valuable information stay behind the data. However, existing re-identification attacks on the AOL and ADULTS datasets have shown that publish such data directly may cause tremendous threads to the individual privacy. Thus, it is urgent to resolve all kinds of re-identification risks by recommending effective de-identification policies to guarantee both privacy and utility of the data. De-identification policies is one of the models that can be used to achieve such requirements, however, the number of de-identification policies is exponentially large due to the broad domain of quasi-identifier attributes. To better control the trade off between data utility and data privacy, skyline computation can be used to select such policies, but it is yet challenging for efficient skyline processing over large number of policies. In this paper, we propose one parallel algorithm called SKY-FILTER-MR, which is based on MapReduce to overcome this challenge by computing skylines over large scale de-identification policies that is represented by bit-strings. To further improve the performance, a novel approximate skyline computation scheme was proposed to prune unqualified policies using the approximately domination relationship. With approximate skyline, the power of filtering in the policy space generation stage was greatly strengthened to effectively decrease the cost of skyline computation over alternative policies.

INTRODUCTION:

IN the age of big data, it is important to exchange and share data among different parties. For example, all registered hospitals in California of US are required to submit specific demographic data on some patients which have been in good condition .However, publishing those data containing sensitive information could violate individual's privacy. In order to get sufficient protection while maintain high data utility, privacy-preserving data publication (PPDP) is becoming an important and interesting research topic.

Not all attributes are sensitive, only some of them which are sensitive need to be protected, like the salary, the name of disease in medical records and so on. Usually, the identification attributes such as id, names and phone numbers are removed from the data table prior to release. However, the published records may still contain quasi-identifiers, such as demographic attributes which contain age, race, gender or post code. Even though the quasi-identifier (*QI*) attributes do not directly reveal individual's identity, but they may appear together with identification attributes in another published datasets, which may lead to linkage attacks to re-identify private information.