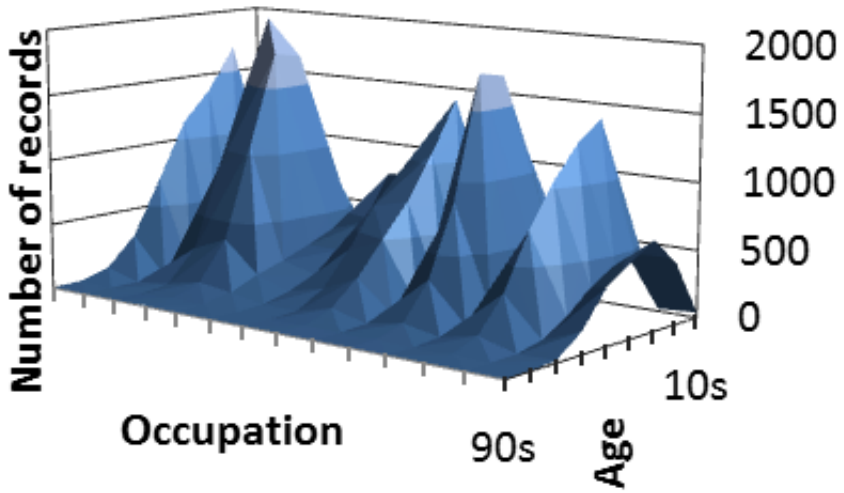
Random additions efficiently anonymize large data sets

December 29 2015



Original (left) and reconstructed anonymized data (right) for age and occupation using the proposed algorithm.

Balancing transparency and freedom of information with the right to privacy lays high demands on data handling methods. So far methods for anonymizing shared data sets have assumed that there is a distinction between details that can be used to identify an individual (quasi-identifiers) and details that are deemed 'sensitive' and private, but this is not always the case. Now Yuichi Sei and Akihiko Ohsuga from the University of Electro- Communications, alongside Takao Takenouchi from NEC Corporation in Japan, have devised an algorithm that

efficiently anonymizes data sets without assuming this distinction.

The researchers use hospital lists as an example. A data set may include the name (direct identifier), address and age (quasi-identifier) and sensitive information (a medical condition). Even without giving the name for each entry, someone using the data set could identify entries from the age and address. In addition, anonymization should be resistant to attempts to identify particulars by comparing two anonymized sets for the same data.

One approach to anonymizing data is to add noise to a data set, where the frequency of each possible value for each attribute is presented in a histogram. However as Sei, Ohsuga and Takenouchi point out this can greatly increase the quantity of the data. "Because almost all of the categories have only a few people in the histogram, the noise added to each category of the histogram has a heavy impact."

The UEC-NEC Corporation researchers instead randomised the data set for each attribute and added random values to each entry. "Through simulations of real data sets, we prove that our proposed method can anonymize and reconstruct databases while keeping a high quality of data within a realistic period." The approach may be useful for anonymizing public records such as the census and electronic electoral votes.

More information: (l1, ..., lq)-diversity for anonymizing sensitive quasi-identifiers 2015 *IEEE Trustcom/BigDataSE/ISPA* 596-603. DOI: 10.1109/Trustcom-BigDataSe-ISPA.2015.424

Provided by University of Electro Communications