Intervals preserve medical information for research but protect patient privacy

G. Xiang[1], J. O'Rawe[1,2], V. Krienovich[3], J. Hajagos[2], and S. Ferson[1]

[1]Applied Biomathematics, Setauket, New York 11733 USA, scott@ramas.com

[2]Stony Brook University, Stony Brook, New York 11794 USA, jazony33@gmail.com

[3]University of Texas at El Paso, El Paso, Texas 79968 USA, vladik@utep.edu

Keywords: *intervals; patient privacy; data informativeness; anonymization;* k-*anonymity.*

Abstract

Data collected during health care delivery and public health surveys possess a great deal of information that could be used in biomedical and epidemiological research. Access to these data, however, is usually restricted because of the private nature of most personal health records. Simple strategies to anonymize data do not protect privacy. For example, dropping explicit identifiers such as name, social security number and address makes the data appear anonymous, but the remaining attributes can often be used to re-identify individuals. Golle (2006) showed that 63% of the US population can be uniquely identified by gender, five-digit ZIP code, and date of birth, and those three attributes often exist in health data put in open archives. This information is widely distributed, so it can easily be obtained and used for re-identification. Techniques for statistical disclosure control have been developed to ensure privacy but information truthfulness is not well preserved so that unreliable results may be released. In generalization-based anonymization approaches, there is information loss due to attribute generalization and existing techniques do not provide sufficient control for maintaining data utility.

We need methods that protect both the privacy of individuals as well as the integrity of the statistical relationships in the data. The problem is that there is an inherent tradeoff between these. Protecting privacy always loses information. However, for a given anonymization strategy, there are often multiple ways of masking the data that meet the disclosure risk criteria provided. This can be exploited to choose the solution that best preserves statistical information while still meeting the disclosure risk criteria. We are developing an integrated software system that provides solutions for managers of data sets so they can minimize disclosure risks while maximizing data informativeness. To overcome the computational challenges associated with subsequent statistical analyses, we selected the shapes of the intervals used in the anonymization to exploit reduced-complexity algorithms available for those analyses.

Reference

Golle, P. Revisiting the uniqueness of simple demographics in the US population. *Proc. of the 5th ACM Workshop on Privacy in Electronic Society*, Alexandria, VA, 2006.