Fri 11/4/2011 4:03 PM

Response to request for comments

As a researcher who conducts research on human subjects, and who is active in training IRBs and others on the ethical issues surrounding human subject research, particularly online, I want to express my concern about the apparently conflicting objectives being pursued by the federal government when it comes to research data associated with humans (whether technically human subjects or not). As has become very clear through highly-publicized examples (the AOL search dataset, the Netflix Prize dataset), and as will become even more clear with any datasets involving personal or medical data, the ability to re-establish links between "deidentified data" and individuals is increasing, and data sets that were once believed to present no risk would today be considered too much at risk for re-identification. There is no reason to believe that this continuing evolution will change.

Part of this challenge is fundamental. It is always easier to seek a match for an individual in a large dataset than to reidentify the full set. Consider a dataset that contains only date of birth, three digits from a social security number, three digits from a phone number, and some sensitive information (e.g., HIV status). From that data set, it would be computationally hard, if not intractable, today to build a list of HIV-positive people. But a person with access to one individual's information could trivially check whether that individual is in the dataset, and find his status if so. This asymmetry, combined with the pace of technological change, is rarely considered adequately in making plans to archive data for permanent use.

Hence, I strongly urge that any policy that addresses mandates for data archival have a very clear section on the policies and practices for opting out in cases where the usefulness of the data cannot be separated from potentially reidentifiable personal information, and for subsequent removal of datasets from archives if unexpected developments lead the datasets to present unacceptable risks to individuals from or about whom the data was collected.

JK

--
--
Joseph A. Konstan
Distinguished McKnight Professor and Distinguished University Teaching Professor
Associate Department Head
Department of Computer Science and Engineering
University of Minnesota