NOTES ABOUT ANONYMIZING DATA FOR PUBLIC RELEASE

ANDREW J. BLUMBERG

Abstract. The purpose of this document is to highlight the difficulties of anonymizing data for public release and advocate a *adversarial challenge* criterion for assessing the threats associated with such release.

1. Overview

The most important principle to keep in mind when thinking about anonyimzing data for public release is that **anonymizing data is extremely difficult**. It is basically *not possible* for non-technical people to assess proposed anonymization techniques and determine the threat (or lack thereof).

There are two kinds of problems that can arise:

(i) **Sparsity:** The Netflix database de-anonymization provides an excellent way of thinking about the problem. The basic observation of this work is that the database of per-individual movie preferences is *sparse* — the average preference set is fairly far from any other preferences. As such, just a little bit of information about an individual's preferences suffices to identify them uniquely. The slogan is *not that many people have your exact preferences*. This appears to be broadly true for many databases with more than about five or ten entries per individual.

(ii) **Quasi-identifiers:** This notion is due to Sweeney, and refers to a small set of fields in a data record that in combination serves to uniquely identify the owner of the record. For instance, knowing a person's ZIP code, gender, and birth date uniquely identifies about 90% of the US population.

(Note that these potential de-anonymization problems are distinct: the Netflix database has no quasi-identifiers, but it is sparse.)

There are a variety of examples of released social science data or medical data being de-anonymized — for instance, released data about Chicago homicides, cross-referenced against the social security death index in a rather unsophisticated analysis, sufficed to identify some 35% of the victims precisely. Anonymized hospital discharge data cross-referenced against voting records was successfully de-anonymized.

In all of these case, the agencies or entities releasing the data made good faith efforts to anonymize the data, and appeared to give some thought to how to do this. The problem is that it's very difficult, and the phenomena that cause trouble (i.e., sparsity) tend to be counter-intuitive.

Date: September 26, 2010.

2. Solutions

Accepting however that it is desirable to attempt to release anonymized data for research purposes, the question arises of what to do. The lesson of the previous discusion is that even good faith efforts at anonymization often fail.

There are some situations in which *provable guarantees* of anonymity exist. For instance, releasing a database consisting only of patient blood type records, for a sufficiently large population, is provable hard to de-anonymize. There are also some systems which exist for supporting anonymization of restricted classes of data. In general, however, formal guarantees are unavailable.

The solution we propose is an *adversarial model*. To assess the threats associated with releasing data, the data should be tested for sparsity and in general a determined attempt should be made to de-anonymize the data. That is, an internal team which does not have access to the original data should use the standard techniques that have emerged for de-anonymization (e.g., the techniques of Narayanan and Shmatikov) to assess the database. Only data which survives this test should be released publicly, and the release should attest to the tests that the data was subjected to.

Department of Mathematics, University of Texas, Austin, TX 78703
E-mail address: blumberg@math.utexas.edu