Institute of Education Sciences
National Center for Education Statistics

# NCES/NISS TECHNICAL REPORT ON DISCLOSURE RISK VS. DATA UTILITY: THE R–U CONFIDENTIALITY MAP

DECEMBER 2001

## EXECUTIVE SUMMARY

Empirical analysis requires access to data. For data about important policy and management issues, information organizations (IOs) - such as government statistical agencies - are the conduit between data providers and data users. However, data confidentiality is a concern for IOs as they work to disseminate products based on collected data that contribute legitimate information to their clients - e.g., government policy makers, individuals, firms, non-governmental organizations, the media, and interest groups. Dissemination can compromise the pledged confidentiality of the data if a *data snooper* - anyone with legitimate access to the data product and whose goals and methods in the use of the data are not consonant with the mission of the agency - is able to gain illegitimate information about a respondent.

Ensuring confidentiality is not a simple task as removal of apparent identifiers like name, social security number, email address, etc. is not adequate to lower disclosure risk to an acceptable level. The key reason being that, today, a data snooper can get inexpensive access to databases with names attached to records. Having this external information, the data snooper can employ linkage techniques, and with such a linkage, the record would be *reidentified*. To publicly disseminate a data product safe from attack by a data snooper, an IO must go beyond deidentification; it must restrict the data by employing a disclosure limitation method.

IOs must manage the not easily resolved tension between ensuring confidentiality and providing access, such that disseminated data products that have both:

1. high data utility U, so faithful in critical ways to the original data (analytically valid data), and
2. low disclosure risk R, so confidentiality is protected (safe data).

This article looks systematically at the simultaneous impact on disclosure risk and data utility of implementing disclosure limitation techniques and choosing their parameter values. This article introduces and exploits the *R-U confidentiality map* and applies it in some important contexts to provide a quantified link between R and U directly through the parameters of the disclosure limitation procedure. With an explicit representation of how the parameters of the disclosure limitation procedure affect R and U, the tradeoff between disclosure risk and data utility is apparent.

**R-U Confidentiality Map**

In its most basic form, an R-U confidentiality map is the set of paired values, (R, U), of disclosure risk and data utility that correspond to various strategies for data release. Typically, these strategies implement a disclosure limitation procedure. Such procedures are determined by parameters, for instance, the magnitude of the error variance $\lambda^2$ for noise addition. As $\lambda^2$ is changed, a curve is mapped in the R-U plane. Visually, the R-U confidentiality map portrays the tradeoff between disclosure risk and data utility as $\lambda^2$ increases, and so more extensive masking is imposed.

Present practice by IOs in assessing tradeoffs between disclosure risk and data utility is primarily heuristic. Recommendation 6.2 of the National Academy of Sciences Panel on Confidentiality and Data Access (Duncan, Jabine and de Wolf 1993) advises the development of theoretical foundations for such determinations. The idea is to view the actors as decision makers, who take actions in light of their perceptions of probabilities and consequences:

1. The data snooper can choose (or not) to make identifications and draw inferences on the basis of the released data product, and
2. The IO chooses a statistical disclosure limitation method to deter the data snooper.

From this perspective, disclosure risk depends on the decision structure - probabilities and utilities of consequences - of the data snooper and IO.

### R-U Confidentiality Map for Additive Noise

The article further demonstrates how an R-U confidentiality map can be constructed to examine the impact of a disclosure limitation procedure that has received substantial attention - additive noise - by considering three different knowledge states, depending on the group to which the data snooper can isolate the target:

1. *Population*. The target $\tau$ has the same distribution as the $X_i$;
2. *Sample*. The target $\tau$ is one of the Xi's, (i.e., is in the sample); or
3. *Record*. The target $\tau$ is not only in the sample, but the data snooper has enough external information to be able to identify (link to) the specific masked record corresponding to $X_i$.

## A Database-Specific Approach: Constructing an Empirical R-U Confidentiality Map

Finally this article demonstrates how an organization can produce and make use of an R-U confidentiality map for a particular database. Using a real-life example of both practical size and realistic complexity, the demonstration details how this *empirical R-U confidentiality map* can be used to:

- inform the IO about whether or not proposed disclosure limitation methods are adequate in lowering disclosure risk and maintaining data utility,
- facilitate comparisons between various disclosure limitation methods, and
- examine the risk of particular types of data snooper knowledge.

Overall, this paper gives a framework for the determination of the parameter values of a disclosure limitation procedure and for the comparison of disclosure limitation procedures. The framework focuses on the tradeoff between data utility and disclosure risk, permitting the agency to consider the tradeoffs between providing more useful data to users and lowering the risk to confidentiality.

Link to the Full Report