

Working Paper No. 8 (Summary)

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (ii): New data release techniques

**DATA MINING METHODS FOR LINKING DATA
COMING FROM SEVERAL SOURCES**

Invited paper

Submitted by University of Catalunya¹

¹ Prepared by Vicenç Torra (vtorra@iia.csic.es); Josep Domingo-Ferrer (jdomingo@etse.urv.es); and Angel Torres (atorres@etse.urv.es).

Data Mining Methods for Linking Data Coming from Several Sources

Vicenç Torra¹ and Josep Domingo-Ferrer² and Àngel Torres²

¹ Institut d'Investigació en Intel·ligència Artificial - CSIC
Campus UAB s/n, E-08193 Bellaterra (Catalunya, Spain)
e-mail vtorra@iia.csic.es

² Dept. Computer Engineering and Maths (ETSE),
Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43006 Tarragona, Catalonia
e-mail {jdomingo,atorres}@etse.urv.es

February 26, 2003

Abstract

Statistical offices are faced with the problem of multiple-database data mining at least for two reasons. On one side, there is a trend to avoid direct collection of data from respondents and use instead administrative data sources to build statistical data; such administrative sources are typically diverse and scattered across several administration level. On the other side, intruders may attempt disclosure of confidential statistical data by using the same approach, *i.e.* by linking whatever databases they can obtain. This paper discusses issues related to multiple-database data mining, with a special focus on a method for linking records across databases which do not share any variables.

Keywords: Statistical disclosure control, Re-identification, Data mining, Artificial intelligence.

1 Introduction

Statistical offices are faced with the problem of multiple-database data mining at least for two reasons:

- On the good side, there is a trend to avoid direct collection of data from respondents and use instead administrative data sources to build statistical data; such administrative sources are typically diverse and scattered across several administration level. Linking administrative information held by municipalities with information held at higher administration levels can yield information that is more accurate and cheaper than the one that would be collected directly from respondents.

- On the bad side, statistical offices must realize that intruders may attempt disclosure of confidential statistical data by using exactly the same approach, *i.e.* by linking whatever databases they can obtain. This is the relevant side for statistical disclosure control (SDC).

This paper discusses issues related to multiple-database data mining, with a special focus on a method for linking records across databases which do not share any variables.

Section 2 is about general concepts of data mining and knowledge discovery in databases. Section 3 discusses the use of data mining in SDC, that is, how data mining can increase disclosure risk.