

Working Paper No. 35 (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 (v): Risk assessment

## **A GRAPH THEORETICAL APPROACH TO RECORD LINKAGE**

**Contributed paper**

Submitted by the Statistisches Bundesamt, Germany<sup>1</sup>

---

<sup>1</sup> Prepared by Rainer Lenz (rainer.lenz@destatis.de).

Item (v) Risk assessment

## A graph theoretical approach to record linkage

Rainer Lenz  
Statistisches Bundesamt  
Gustav-Stresemann-Ring 11  
65189 Wiesbaden, Germany  
rainer.lenz@destatis.de

### Extended abstract of supporting paper

#### 1. Introduction

The software package  $\mu$ -ARGUS offers a variety of methods to produce Safe micro-data files. Users of the package are offered a choice between methods, or they must select suitable parameters when applying a method to a data set. While a particular method, or setting of parameters, may perform well with respect to information loss, it may not protect the data sufficiently according to the user's requirement. In order to assess the performance of a disclosure limitation method, it is not enough to consider its behaviour regarding the loss of information alone. The disclosure risk associated to the protected data set must also be taken into account. Record linkage is a practical disclosure risk assessment methodology, applicable to every masking method offered by the  $\mu$ -ARGUS package. It has therefore been decided to add a record linkage tool to the package. While work on development of this tool is still in an early stage, this paper will describe the approach currently foreseen for the implementation. The paper will conclude with a small, illustrative application of record linkage methods to data of the german structure of costs survey.

#### 2. Description of the method

Let  $A$  and  $B$  be two data files of observations, corresponding to subsets of a given set of objects. Both files contain records of variables and share a nonempty set  $v_1, \dots, v_k$  of common variables. The aim is to decide for each pair  $(a, b)$  of records taken from  $A \times B$  whether  $a$  and  $b$  belong to the same object. The paper will describe a 3-phase algorithm of polynomial run-time, as outlined below.

Throughout the paper we will use the following denotation: a **graph**  $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$  is a relational structure, consisting of a set  $V(\mathcal{G})$  of vertices and a set  $E(\mathcal{G}) \subseteq V(\mathcal{G})^2$  of edges.  $\mathcal{S} = (V(\mathcal{S}), E(\mathcal{S}))$  is called **subgraph** of  $\mathcal{G}$ , if it holds  $V(\mathcal{S}) \subseteq V(\mathcal{G})$  and  $E(\mathcal{S}) \subseteq E(\mathcal{G})$ . It are considered undirected graphs,

fulfilling  $\forall(a, b) : (a, b) \in E(\mathcal{G}) \Rightarrow (b, a) \in E(\mathcal{G})$ . That is,  $E(\mathcal{G})$  determines a symmetric binary relation. The edge  $(x, y) \in E(\mathcal{G})$  is said to be **incident** with the vertices  $x$  and  $y$ , and  $x, y$  to be **adjacent**.  $\mathcal{G}$  is called **bipartite graph** with bipartition  $(X, Y)$ , if  $V(\mathcal{G})$  is a disjoint union  $V = X \cup Y$ , such that every edge  $e$  is incident with both an  $x \in X$  and an  $y \in Y$ . Moreover, if every  $x \in X$  is connected with every  $y \in Y$ , the graph  $\mathcal{G}$  is said to be **complete**.

In the first phase of the algorithm, for each pair  $(a, b)$  of records distances  $d_i(a^{(i)}, b^{(i)})$  are calculated. The distances are used to build up a complete vector-weighted bipartite graph. In phase 2, the resulting multiobjective linear program is linearly extended and transformed into a single objective assignment problem with weights  $w_{ij} \geq 0$  :

$$\text{Minimize } \sum_{i=1}^n \sum_{j=1}^n w_{ij} x_{ij}, \quad (\mathbf{AP})$$

$$\text{s.t. } x_{ij} \in \{0, 1\} \quad \text{for } i, j = 1, \dots, n,$$

$$\sum_j x_{ij} = 1 \quad \text{for } i = 1, \dots, n \quad \text{and}$$

$$\sum_i x_{ij} = 1 \quad \text{for } j = 1, \dots, n.$$

Let a **matching**  $\mathcal{M}$  of  $\mathcal{G}$  be a subgraph with the property that no two edges are adjacent in  $\mathcal{M}$ . If  $v \in V(\mathcal{M})$ , then  $\mathcal{M}$  **saturates**  $v$ . Moreover, if every  $v \in V(\mathcal{G})$  is saturated,  $\mathcal{M}$  is called **perfect matching**.

Using this terminology, problem **(AP)** can be interpreted as follows: Let  $\mathcal{G}$  be a complete weighted bipartite graph with bipartition  $(A, B)$ , where  $w_{ij} = w(a_i, b_j)$  are the weights of the edges  $(a_i, b_j)$ . The constraints enforce a subgraph connecting every  $a_i$  with exactly one  $b_j$  and vice versa (i.e.  $x_{ij} = 1$  if and only if  $a_i$  is connected with  $b_j$ ). Hence, one has to determine a perfect matching of minimal weight and **(AP)** reduces to

$$\text{Minimize } \sum_{(a_i, b_j) \in E(\mathcal{M})} w_{ij}, \quad (\mathbf{AP}')$$

$$\text{s.t. } \mathcal{M} \text{ is matching of } \mathcal{G}.$$

The core of the algorithm is to find an optimal assignment. The idea for phase 3 is to use for this purpose two greedy heuristics and a modification of Kuhn's algorithm<sup>1</sup>. The paper will conclude with an illustrative application of files of real life business micro data – original versus masked data.

<sup>1</sup>H.W.Kuhn, "The Hungarian method for the assignment problem", *Naval Res. Logist. Quart.* 2 (1955), pp. 83-97.