Disclosure Risk of Anonymized Business Microdata Files - Illustrated with Empirical Key Variables

Dr. Daniel Vorgrimler
Federal Statistical Office of Germany
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany
daniel.vorgrimler@destatis.de

Dr. Rainer Lenz
Federal Statistical Office of Germany
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany
rainer.lenz@destatis.de

1 Introduction

Regarding a particular anonymization method, two aspects have to be taken into account: data protection and information loss (see Sturm, 2002). In this paper we study the disclosure risk associated to a confidential business data file (target data). We distinguish between two types of data attack. On the one hand `Database cross match` and on the other hand `Match for a single individual` (see Elliot/Dale, 1999). To disclose a data set the "data intruder" needs additional information – e.g. an outside database – containing a certain number of identical variables (`key` or `matching variables`) with the target data.

Within a database cross match attack, the data intruder matches an outside database with the whole target data set. In order to enhance his outside database, he tries to assign as many true pairs of records as possible.

The intention behind the single individual match is to gain information about a specific target individual. The data intruder collects information about the target individual, using several sources of information. The collected information is then used to identify the target individual, in order to get additional information about it.

2 Database cross match

In the following let without loss of generality the number of records in the outside database equal the number of records in the target database and let $\{v_1, \ldots, v_k\}$ be a nonempty set of key variables.

We present a distance based method, where the distance function d maps from the cartesian product of both databases into the set of non-negative real numbers. For each record pair r and for each variable v_i it is determined a distance $d_i(r)$. Let $\mathcal{A}$ denote a $(1-1)$-assignment matching every record of the outside database uniquely with one of the target data base. Since there is in general no assignment $\mathcal{A}$ which minimizes the sum $d_i(\mathcal{A}) := \sum_{r \in \mathcal{A}} d_i(r)$ for all $i = 1, \ldots, k$ simultaneously, we turn over to a parametric optimization problem (as in Schweigert, 1995) by using weighted sums

$$\lambda_1 d_1(\mathcal{A}) + \ldots + \lambda_k d_k(\mathcal{A})$$

for the k objectives $d_1, \ldots, d_k$, where $\lambda_i > 0$ for all $i = 1, \ldots, k$ and $\sum_{i=1}^{k} \lambda_i = 1$. The choice of the parameters $\lambda_1, \ldots, \lambda_k$ depends on the individual ranking of key variables and the specific knowledge of the decision maker. To solve the resulting single objective linear program, we go back to classical techniques of optimization theory like the most common simplex method.

3 Match for a single individual

A data intruder has personal information and response knowledge about a specific target record. Furthermore he can generate additional information about commercial databases and generally accessible information (e.g. annual reports of enterprises). With this knowledge he tries to identify the target record within a business survey and to disclose information about the corresponding enterprise.

In the "German Structure of Costs Survey" such a data attack was simulated. We repeated the single individual match for 41 enterprises, without consideration of commercial databases (for more information see Vorgrimler, 2003). Note that the dataset was only weakly anonymized. The most important key variables were the regional label (with 9 categories), the business classification and the number of employees. Further keys were total revenue, research and development investments (yes or no), trade activity (yes or no) and the number of active owners. Note that the keys were not available in every case. With these keys 19 of the 41 enterprises could be identified. It was found out that the probability to identify an enterprise increases with the size of the enterprise. Thus, only one record could be identified among the 15 enterprises with less than 250 employees. On the other hand, among the larger enterprises a total of 18 out of 26 could be identified. Faults in the additional information were main reasons for 22 unsuccsessful attempts. The knowledge was particularly inaccurate for the key variables "business classification", "employees" and "total revenue".

4 Conclusion

The identification risk for small enterprises turns out to be low, while for larger enterprises the authors recommend further anonymization to decrease the probability of identification. In general, the target database is partially protected by erroneous values in the empirical key variables.

REFERENCES

Elliot, M., Dale, A. (1999). Scenarios of attack: the data intruder's perspective on statistical disclosure risk. Netherlands Official Statistics, pp. 6-10.

Lenz, R. (2003). A graph theoretical approach to record linkage. Joint UNECE/Eurostat work session on statistical data confidentiality, Luxemburg.

Schweigert, D. (1995). Vector Weighted Matchings. Combinatorial Advances (eds C.J.Colbourn, E.S.Mahmoodian), Kluwer, pp. 267-276.

Sturm, R. (2002). Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten. Allgemeines Statistisches Archiv, Vol. 86, pp. 468-477.

Vorgrimler, D. (2003). Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios, appears in Forum der Bundesstatistik Vol. 40.

Winkler, W.E. (1994). Advanced Methods for Record Linkage. Statistical Research Report Series. Bureau of the Census, Washington DC.

RÉSUMÉ

Un pirate informatique tente, à l'aide d'informations complémentaires extérieures, de réidentifier des ensembles de données d'entreprises dans un fichier protégé. Les auteurs nous indiquent deux voies a suivre. Lors d'un coup de filet général, comme on l'appelle, le pirate compare une banque de données externe avec le fichier protégé, dans le but d'obtenir le plus grand nombre de classements corrects. Lors de l'attaque ciblée, le pirate récolte sur une entreprise particulière le plus grand nombre d'informations possible par lesquelles il tente de réidentifier l'entreprise dans le fichier protégé.