WP.3
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 UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

MICRODATA SHARING VIA PSEUDONYMIZATION

Invited Paper

Prepared by David Galindo, University of Malaga, Spain and
Eric R. Verheul, Radboud University Nijmegen, PricewaterhouseCoopers Advisory, Netherlands

Microdata sharing via pseudonymization

David Galindo*, Eric R. Verheul**,***

* Department of Computer Science, University of Malaga, Spain. (dgalindo@lcc.uma.es)

** Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands. (eric.verheul@cs.ru.nl)

*** PricewaterhouseCoopers Advisory, The Netherlands. (eric.verheul@nl.pwc.com)

Abstract. Individual data records are essential for empirical research, and yet due to the very precious information they contain, their release poses a problem to the confidentiality of the individuals concerned. In this paper we give a high level description of a privacy-preserving microdata sharing system wherein subjects identifiers are replaced by cryptographic pseudonyms. The resulting system facilitates information sharing between organizations that typically are not allowed to exchange the microdata they own.

1 Introduction

Individual data records are essential for empirical research, and yet due to the very precious information they contain, their direct release thwarts the confidentiality of the individuals concerned. The fact that research is interested in collective features rather than individual distinctiveness, makes it possible to reconcile data utility and individual confidentiality: data identifiers can be removed or encoded and data fields can be modified by means of statistical disclosure controls, while overall the collective features of the resulting de-identified data are preserved.

Microdata comes from heterogenous sources, such as statistical offices, hospitals or insurance companies to name a few. There are a number of parties, named as Researchers, interested in getting access to this data for economical or research purposes. In the case of national statistical offices, Researchers face in general two modes of accesses: either access to the microdata is granted in the premises of the national statistic authorities; or the microdata is anonymized and released to Researchers under certain conditions. In both cases, the original data has been modified to preclude the direct identification of the subjects.

The aim of this paper is to describe privacy-preserving microdata sharing systems obtained by replacing subjects identifiers with pseudonyms with special mathematical and cryptographic properties. The pseudonymizing system is controlled by a Trusted Third Party (TTP), and no party in the scheme (except the TTP) can re-identify the individuals from the pseudonyms. Still, natural set operations between

different pseudonymized databases, like database union and intersection are supported. These operations allow for flexible research of personal data of individuals residing at different organizations that typically do not share information.

2 Pseudonymous data sharing

Consider a *database* consisting of entries of the form $(id, D(id))$, where id is the identifier field (also called identity) and $D(id)$ is the data field. A *pseudonymized database* is obtained by replacing the identity id in the database entries by a blinded identifier $P(id, O)$, called *pseudonym*. The blinded identifier $P(id, O)$ does ideally not leak any information on the identity id. The individual with identity id is only known to the Organization O by its pseudonym $P(id, O)$, and the key property is that the organization O is not able to link together $P(id, O)$ and id (under certain cryptographic assumptions). This property is called *pseudonymity*.

Pseudonymized databases with the above properties provide a virtually unexplored tool for building privacy preserving information sharing systems. Roughly speaking, an information sharing system is called *privacy-preserving* if no information is leaked on individuals identities. We stress the latter is interpreted from a strict cryptographic point of view, that is, the qualification privacy-preserving refers to the cryptographic techniques used for pseudonymization and related operations, since from a global point of view privacy-preserving pseudonymized microdata sharing systems likely do not exist. The reason is simple, even though the data is pseudonymized, there is the risk that the characteristics of the data singles out a person, e.g. by a combination of profession, age and place of residence. This risk of *indirect identification*, cf. [6, 3], becomes even larger when linking several pseudonymized databases, which is one of our targets. The issue of indirect identification is outside the scope of this paper and is covered by an abundant literature [1]. Although out of scope, it is our position that indirect identification should be an important point of attention in deciding which data Researchers are provided with; at the very least a Researcher should only get the information required for his Research and nothing more.

An example is illustrative. Suppose a Researcher wants to find out the correlation between certain pharmacy usage and traffic accidents, e.g. with the objective to provide for better warnings on the usage of certain pharmacy in traffic. However privacy laws prevent the Suppliers holding the data from releasing this information. Let us assume that the representation of identities of individuals is unique, e.g. takes the form of a Social Security Number. We can achieve the Researcher's desired functionality while circumventing the Suppliers concerns by providing the Researcher with two kinds of data: pseudonymized drug usage of individuals from pharmacies

[1] The interested reader is referred to [6] for an introduction to this topic, and to [7] for a state of the art.

and pseudonymized traffic accidents data from insurance companies.

With the pseudonymized data received, the Researcher can easily compute its target correlation, since an individual that occurs in the non-pseudonymized databases of pharmacies and insurers leads to the same pseudonym $P(id, R)$ in the Researcher's database. We name this information sharing technique as *pseudonymous data sharing*. Obviously, a malicious Researcher is tempted to learn the identities of the individuals involved. Still, a misbehaving Researcher is prevented from learning information on the identities thanks to the use of pseudonyms. Additionally and depending on the application, one might require that two Researchers R_i and R_j should not be able to match their pseudonymized databases, and thus the pseudonyms $P(id, R_i)$ and $P(id, R_j)$ must necessarily be different and unlinkable.

This in principle complies with the Recommendation 83 (10) of the Council of Europe [10, 12] on the protection of personal data collected and processed for statistical purposes, since the recommendation states that

> An individual should not be regarded as 'identifiable' if the identification requires an unreasonable amount of time, cost and manpower.

3 Applications

There are several contexts where this technique is useful. Consider the case of several organizations working with different responsibilities on the same set of (potential) individuals. These responsibilities prevent those organizations from sharing information unless certain concurrences occur. Alas, one cannot detect concurrences without sharing information, thus ending in a catch-22 situation. By sharing on a pseudonymous basis one can break through this situation. This catch-22 situation does not come without consequences. Indeed, it was one of the reasons the 9/11 tragedy was not timely foreseen; the different agencies (law enforcement, secret service) did not share their information in a timely fashion, see the 9/11 report [5, p.416]. It appears that this situation is also present in the context of child-abuse. According to [4], if social workers, nurses, midwifes, psychiatrists, police share their information in a timely fashion then serious child abuse can be prevented. In both situations one can use our techniques to develop a Reference Registry containing all individuals on a pseudonymous basis; as soon as a concurrence occur, a special procedure is taken to identify the responsible individual.

Other applications include the integration of separated clinical and biological data (e.g. genomic data) in common databases [2, 13]. In case relevant diagnostic findings are revealed in the join of the databases, only a trusted party (to be introduced later) can re-identify the patient from the pseudonym. If the results gained in the research can positively influence the patient's therapy, the patient can be contacted and be told the results. This and the former example show that

pseudonymization has advantages over anonymization, since pseudonyms enable re-identification by a trusted authority, whereas anonymization does not.

4 Our framework

The framework for pseudonymized data sharing we propose is based on three main considerations. In the first place, we observe that the main task of Supplier organizations is very often different from that of supplying information to Researchers. Following our example, this would be the case for pharmacies or insurers, while a governmental agency taking care of traffic matters falls probably outside of this category. In our view it is important to keep Suppliers workload as low as possible in a practical information sharing system. To this aim we introduce in our model an intermediary organization called *Accumulator*. These are buffers between Suppliers and Researchers which are responsible for keeping pseudonymized versions of Suppliers databases and feeding Researchers with data. From time to time, Suppliers hand over their data in pseudonymized form to one or more Accumulators that each have their own sets of pseudonyms $P(id, A_i)$. Researchers are supplied pseudonymized data by Accumulators under pseudonyms $P(id, R_j)$. Accumulators exchange data in a sensible manner too. Suppliers have their workload alleviated since it is supposed they hand over data to Accumulators rarely, for instance only when important database updates need to reflected. In contrast, Researchers will ask Accumulators for data much more frequently.

Secondly, the allowance of these protocols and the type of data that is sent along with the protocols is governed by the Regulatory Privacy Body (RPB) from a functional perspective. We envision that a strict licensing infrastructure for protocols will be enforced by the RPB, describing:

1. Which parties are allowed to perform what protocols with each other.
2. What kind of data may be sent along with the protocols.
3. What kind of subsets of identities (Suppliers) or pseudonyms (Accumulators) are allowed as input to the protocols.

From a technical perspective the execution of all protocols depends on cryptographic keys governed by a Trusted Third Party (TTP). We want the involvement of this TTP to be very low, at least from a computational and availability point of view.

Thirdly, we deliberately choose to not support protocols between Researchers as they are assumed malicious and might deviate from protocols, e.g., by sending along disallowed data with the pseudonyms.

Supported operations. Let us now outline the operations enabled by a pseudonymous data sharing system. The execution of all protocols requires cryptographic

keys governed by a Trusted Third Party (TTP); the parties in our system need to be handed in advance certain cryptographic keys by the TTP (see Figure 1).

Supplier Pseudonymization The result of the operation $S_i \rightarrow_P A_j$ is that a (selected) list of identities id in Supplier S_i database is provided to Accumulator A_j under his pseudonyms $P(id, A_j)$. This operation typically needs to be performed periodically to reflect updates of the Supplier's database. By sending along data related to these identities, Accumulator A_j gets a pseudonymized version of S_i's database (or part thereof). After this operation has been performed, Supplier S_i can not link together id and $P(id, A_j)$.

Accumulator Exchange The operations $\rightarrow_\cup$ and $\rightarrow_\cap$ (pseudonyms union and intersection respectively) enable two Accumulators to perform the two atomic set operations union and intersection on their pseudonymized data collections.

operation $\rightarrow_\cup$: After the operation $A_i \rightarrow_\cup A_j$ the Accumulator A_j possesses pseudonyms of the form $P(id, A_j)$ for every id that was available in pseudonymized form in either A_i's or A_j's database. By sending along data related to these identities, Accumulator A_j gets a pseudonymized join with A_i's database.

operation $\rightarrow_\cap$: After the operation $A_i \rightarrow_\cap A_j$ the Accumulator A_j knows which of his pseudonyms also occur (as peers) in Accumulator A_i's database. If A_i relates his data to these pseudonyms then a specific encryption feature in the $A_i \rightarrow_\cap A_j$ operation facilitates that A_j can only decrypt the data for pseudonyms in the intersection.

Researcher Provisioning As the result of the data provisioning operation $A_i \dashrightarrow_D R_j$ between an Accumulator A_i and Researcher R_j a list of pseudonyms in an Accumulator's database is provided to the Researcher under his pseudonyms $P(id, R_j)$. By sending along data related to these pseudonyms, Researcher R_j gets a pseudonymized version of A_i's database.

5 Security properties

The TTP is assumed honest, i.e. it will not deviate from protocols or try to deduce secret information from the information it gets as part of protocol execution. Suppliers and Accumulators are assumed honest but curious. That is, they will not deviate from protocols but might try to deduce secret information from the information they get. In practice one should try to ensure that Suppliers and Accumulators are honest; the fact that our model allows them to be curious should be seen as an indication that there is some "room for error". Finally and most importantly, we assume that Researchers are malicious. They are willing to deviate from protocols

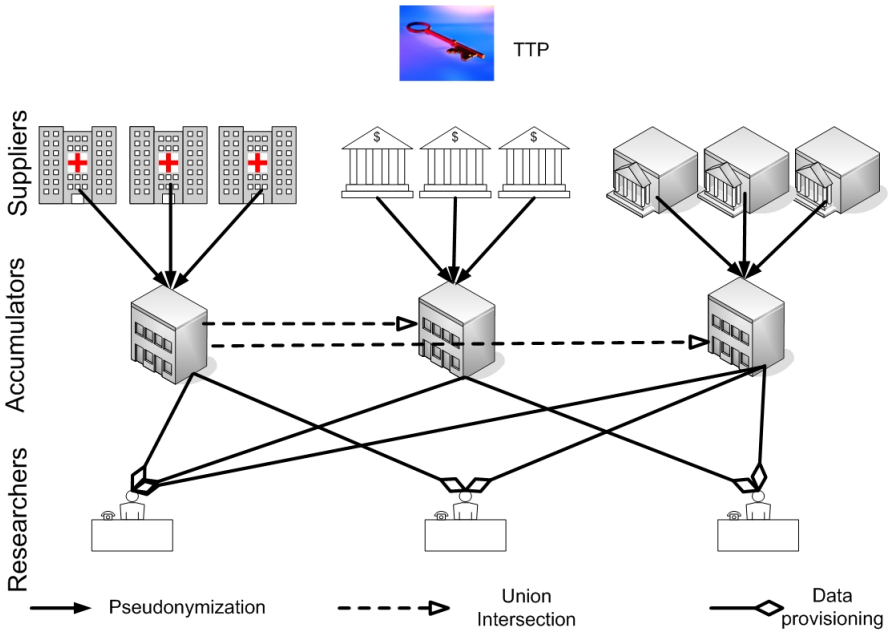

Figure 1: Overview

and to share their cryptographic keys with other Researchers in order to deduce secret information; most notably to relate pseudonyms of other parties or to realize pseudonymity removal.

We require two basic security properties. First and foremost is the prevention of pseudonymity removal: no agent in the scheme (Suppliers, Accumulators or Researchers) should be able to relate through cryptanalysis a given pseudonym with a given identity even if this party was able to do so for many pseudonyms by other means (most notably through indirect identification). This property is called *pseudonymity*.

The second property deals with the prevention of pseudonym matching of different parties. If a party X in the scheme gets hold of a copy of a database of another party Y, then party X should not be able to relate his database with that of Y without the collaboration of Y, even if X and Y previously run the pseudonymized intersection protocol and X is able to relate the databases partially by non-cryptographic means. Additionally, the sender in a protocol should not be able from engaging in the protocol to deduce information about which pseudonyms are already in the receiving party database. These properties are named as *mutual separation*. This property considers the scenario where X's pseudonymized database is renewed periodically, and where the fact that Y was allowed to match its database to that of X does not imply this matching is allowed at a later time. Notice that the mutual separation property excludes the possibility that the pseudonymous data sharing operations are

implemented by the parties through a transition table between pseudonyms, that is, a table of pseudonym pairs $\{(P(id, A_s), P(id, A_d))\}$ or $\{(P(id, A_s), P(id, R_u))\}$.

6 Designing a pseudonymous data sharing system

We start by outlining a straightforward implementation of pseudonymous data sharing system based on symmetric encryption. This implementation enjoys high efficiency due to the use of symmetric encryption, but suffers from requiring an on-line TTP to perform pseudonymization and database operations, as well as scalability and flexibility drawbacks.

In this implementation, the TTP selects a blockcipher $\big(\mathrm{Enc}_{K_i}(\cdot), \mathrm{Dec}_{K_i}(\cdot)\big)$ and generates for each Accumulator A_i a secret key K_i which is unknown to A_i. The database of Supplier S_j is denoted as database rows $\big(id, D(id)\big)$. Supplier S_j sends the datablocks $D(id)$ directly to A_i, i.e. the identity field is removed. The identities id are sent (in the same order as the datablocks were sent to A_i) to the TTP. The TTP encrypts the identities using K_i, leading to $P(id, R_i) = \mathrm{Enc}_{K_i}(id)$ and sends them to A_i; these constitute A_i's pseudonyms. The Accumulator A_i joins the information received from the Supplier and the TTP in the same order, leading to the pseudonymized database $\big(P(id, A_i), D(id, i)\big)$. The union operation $A_i \rightarrow_{\cup} A_m$ would be as follows: Accumulator A_i sends the data blocks $D(id, i)$ of his choosing directly to A_m and sends his pseudonyms to the TTP in the same order. On receipt the TTP does a decrypt operation $id = \mathrm{Dec}_{K_i}\big(P(id, A_i)\big)$ and an encrypt operation $\mathrm{Enc}_{K_m}(id)$ (leading to the pseudonyms of A_m) and sends these to A_m in the same order as received. On receipt R_m joins the information received from the Researcher R_m and the TTP in the same order, leading to the pseudonymized database $\big(P(id, A_m), D(id, i)\big)$. Finally, R_m joins the latter with its own database $\big(P(id, A_m), D(id, m)\big)$. The rest of the operations are implemented analogously.

This implementation enjoys efficiency and high security properties, but requires an on-line TTP, which is a burden to the system. It turns out that an efficient pseudonymous data sharing system with low involvement of the TTP (only in charge of distributing at once cryptographic keys) is possible by using more sophisticated asymmetric cryptography techniques. The details can be found in [9].

7 Related work

Although the framework of our pseudonymous data sharing scheme is actually found in practice [13, 3], it seems that it has not received much attention in the cryptographic literature. What has received considerable attention is the case in which the holders of the private databases (Suppliers in our framework) are themselves interested in sharing information while preserving privacy. Examples of that are the data mining protocols using secure multiparty computation in [11], or works like [1, 14, 8] focusing on concrete operations like database union or intersection. However, none

of these approaches seem to be directly applicable to our framework. Applying these tools in our setting would imply that if a Researcher R_j wants to perform operations on the databases owned by S_l and S_t, these Suppliers must run themselves the protocol and deliver the results to the Researcher. This is not acceptable here, as we want to avoid Suppliers doing the bulk of the work. Additionally, we would like to note that none of these works provide non-interactive protocols for implementing database intersection, in contrast to our concrete scheme [9].

References

[1] Rakesh Agrawal, Alexandre V. Evfimievski, and Ramakrishnan Srikant, *Information sharing across private databases.*, SIGMOD Conference (Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, eds.), ACM, 2003, pp. 86–97.

[2] Russ B. Altman and Teri E. Klein, *Challenges for biomedical informatics and pharmacogenomics*, Annual Review of Pharmacology and Toxicology **42** (2002), 113–133.

[3] Dutch Data Protection Authority, *Landelijke zorgregistraties (national healthcare registrations)*, www.dutchdpa.nl, 2005.

[4] Marian Brandon, Jane Dodsworth, and Daphne Rumball, *Serious case reviews: learning to use expertise*, 2005, pp. 160–176.

[5] 9/11 commission, *The 9/11 report*, www.9-11commission.gov/report/911Report.pdf, 2004.

[6] Josep Domingo-Ferrer (ed.), *Inference control in statistical databases, from theory to practice*, Lecture Notes in Computer Science, vol. 2316, Springer, 2002.

[7] Josep Domingo-Ferrer and Luisa Franconi (eds.), *Privacy in statistical databases, cenex-sdc project international conference, PSD 2006, rome, italy, december 13-15, 2006, proceedings*, Lecture Notes in Computer Science, vol. 4302, Springer, 2006.

[8] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas, *Efficient private matching and set intersection*, Advances in Cryptology - EUROCRYPT 2004, Lecture Notes in Computer Science, vol. 3027, Springer, 2004, pp. 1–19.

[9] David Galindo and Eric R. Verheul, *Pseudonymous data sharing*, 2007, Manuscript.

[10] Franz Kraus, *Data protection and access to official microdata for european research*, NESSIE SRoundtable 4 Access to Quality Comparative Data for European Comparative Socio-Economic Research, 2004.

[11] Yehuda Lindell and Benny Pinkas, *Privacy preserving data mining.*, J. Cryptology **15** (2002), no. 3, 177–206.

[12] Council of Europe, *Recommendation no.r(97) 18 on the protection of personal data collected and processed for statistical purposes*, 1997.

[13] Klaus Pommerening and Michael Reng, *Medical and care compunetics 1*, ch. Secondary Use of the EHR via Pseudonymisation, pp. 441–446, IOS Press, 2004.

[14] Alberto Maria Segre, Andrew Wildenberg, Veronica Vieland, and Ying Zhang, *Privacy-preserving data set union.*, Privacy in Statistical Databases, 2006, pp. 266–276.