Secure big data collection and processing: Framework, means and opportunities

Li-Chun Zhang[1,2,3] and Gustav Haraldsen[2]

[1]University of Southampton (L.Zhang@soton.ac.uk)
[2]Statistics Norway
[3]University of Oslo

Abstract

Statistical disclosure control is important for the dissemination of statistical outputs. There is an increasing need for greater confidentiality protection during data collection and processing by National Statistical Offices. In particular, various transactions and remote sensing signals are examples of useful but very detailed big data that can be highly sensitive. Moreover, possible conflicts of interest may arise for data suppliers who operate commercially. In this paper, we formulate statistical disclosure control for data collection and processing as an optimisation problem. Even when it is difficult to specify and solve the problem unequivocally, the formulation can still provide the basis for comparing different disclosure control methods. We develop a general compartmented system that adapts and implements non-perturbative methods in the related fields of linking sensitive data and secure computation. We illustrate how the system can be configured to yield variously required tables and microdata sets with sufficiently low disclosure risks.

Keywords: Non-survey big data, statistical disclosure control, confidentiality protection, trusted execution environment

1 Introduction

"Survey respondents are usually provided with an assurance that their responses will be treated confidentially. These assurances may relate to the way their responses will be handled within the agency conducting the survey or they may relate to the nature of the statistical outputs of the survey..." (Skinner, 2009). To protect confidentiality, statistical disclosure control (SDC), also known as statistical disclosure limitation (SDL), has long been an important topic when it comes to the *dissemination* of statistics in the form of tables, queries or microdata sets. Many techniques for assessing identification risks and methods for controlling (or limiting) these risks have been developed. See e.g. Elliot and Domingo Ferrer (2018) for a recent overview.

Meanwhile, there is an increasing need for greater protection of confidentiality as the National Statistical Office (NSO) collects and processes big data, en route to the statistical outputs. On the one hand, the ever-more digitalised life form has created a multitude of non-survey big data sources, offering potentially many opportunities for better, quicker and richer statistical outputs, which would have been either extremely demanding or simply infeasible based on traditional surveys. On the other hand, the regulations for protecting confidentiality are being strengthened, and public critical awareness is rising (e.g. Zuboff, 2019) against the detrimental effects when the confidentiality of individual's (or other data subject's) data is breached. In particular, the EU's General Data Protection Regulation (GDPR) requires that businesses and agencies that handle personal data must implement measures to safeguard the data, which could raise barriers for an NSO which would like to use such data for statistical purposes.

The projects of ESSnet Big Data II (2020) organised by Eurostat provide examples of big data for official statistics. The legacy projects (from ESSnet Big Data I) are Online job vacancies and enterprise characteristics, Electricity smart meters, Maritime Automatic Identification System (AIS) and Mobile phone data. The sources being piloted in ESSnet Big Data II are Financial transactions (of many types) and Earth Observation (EO), especially the Sentinels of the EU Copernicus Programme. Table 1 groups the most important *non-survey* big data sources we have in mind for this paper into four *types*: (i) administrative registers, or simply *Register*, (ii) various financial transactions, or simply *Transaction*, (ii) *Remote sensing*, where the sensors can be either fixed or mobile, and (iv) *Internet*. In particular, many new sources of transactions and remote sensing signals obviously have great potential for official statistics.

Table 1: Non-survey big data sources by type.

Type of source	Example of data
Register	vital events, diagnoses
	wage, income tax, VAT, welfare payments
Transaction	scanner data price, point-of-sales receipt
	bankcard or giro payment
	B2B or B2P invoice
	property sales contracts, ownership registration
Remote sensing, fixed	smart meter readings
	weather station readings
	traffic loop signals
Remote sensing, mobile	satellite images, drone images
	airborne laser scanning
	maritime AIS, lorry tracking signals
	mobile phone signals
Internet	web pages
	social media posts

When collecting and processing sensitive data from surveys or administrative sources, the NSO applies *pseudonymisation* as the standard practice regulated by the GDPR,

whereby direct identifiers such as Person Identification Number or Name-Surname are replaced by a *master key* that exists only within the NSO. The data including the master key are encrypted before storage. The information that allows pseudonymisation to be reversed is kept separately from the data. Moreover, any statistical outputs produced on the basis of these data are subject to SDC treatment before dissemination.

Despite the protection provided by pseudonymisation and output disclosure control, our experiences suggest that confidentiality related issues can contribute to either stop or considerably slow down the development of many new big data sources residing with private companies or commercial operators. The data in these sources can typically be much more detailed than in traditional sources and, hence, potentially much more sensitive. For instance, mobile phone locations are much more detailed than any travel data that can be collected by questioning survey respondents, and detailed locations can obviously be personally sensitive. An equally important issue that often arises is when the data exist in a number of competing businesses. For instance, to acquire purchase transactions data from a supermarket chain, but not its competitors, may easily create conflicts of interest and cause reluctance to comply. For the sake of brevity here, we do not go into the many other confidentiality related issues we have experienced, but simply notice that any of them can easily lead to a lengthy process, whereby the NSO needs to justify its request and negotiate compliance.

We argue that offering greater protection of confidentiality (beyond pseudonymisation) *during data collection and processing* can help to alleviate the pressure, making it easier to gain trust and acceptance from data owners, stakeholders and the public, and thereby smooth the access to useful new big data sources. Specifically, in the remainder of this paper, we shall focus on three key aspects of an approach to secure data collection and processing: *conceptual framework, means* and *opportunities*. The discussion will draw on related fields of so-called privacy-preserving computation and data linkage techniques (e.g. PPT Task Team, 2019; Christen et al., 2020).

In Section 2, we describe the concepts central to our approach. Our usage of the terminology will be clarified, where difference and ambiguity exist between SDC and the related fields. We then propose a general protocol governing secure data collection and processing. Finally, we shall formulate SDC for secure big data collection and processing as an optimisation problem, which differs to the formulation for dissemination of statistical outputs given by Skinner (2009).

Given the conceptual framework, we develop a *compartmented system* in Section 3. Configuring the system in various ways to suit different situations can provide the means to greatly reduce the disclosure risks during data collection and processing. The design is inspired by the relevant ideas underlying the so-called *trusted execution environment* (TEE, GlobalPlatform, 2011) which, narrowly speaking, refers to a secure area (enclave) of the main processor of a device, such as a smart phone, whose memory or execution state is invisible to any other process, including the device's operating system. Sabt et al. (2015) point to the separation kernel as a fundamental element in all TEE implementations. Similarly, on entry to the compartmented system, any data set is divided into separate packages of unit IDs and attributes, and the processes that are most critical to disclosure

risks are isolated from one another and inaccessible to the process owner.

Instead of merely considering the undertakings required of secure data collection and processing as extra troubles, we prefer to welcome them as the means for providing opportunities to better, quicker and richer data that can both improve and enlarge the outputs of official statistics. Unlike (sample or census) survey data that arise from probing the respondents for the required information, many non-survey data (such as Transactions) derive their content directly from automatic digital records. Such a *content-orientated* approach can have advantages compared to *unit-orientated* surveys, provided the target measurement is factual and the digital records can form a reliable basis of the responses that one ideally could have obtained by surveying the subjects. For instance, purchase transactions can give more accurate measures of the actual expenditure on the corresponding occasions than those based on diary reports by the sampled household. Not only does this remove the survey response burden, it can also be quicker to adapt the information needs than changing and implementing the designed survey instruments.

In Section 4, we outline several applications where remote sensing and transactions data can either replace or enhance existing sample surveys, or provide new statistics that are either infeasible or impractical via surveys. Of course, non-survey big data can have their own challenges regarding linkage, coverage (or selection) and measurement, such that statistical adjustments and uncertainty assessment are generally required. Although secure data collection and processing neither can nor is intended to solve all these problems directly, it is often a critical and necessary part of solutions to using non-survey big data for official statistics.

Some final remarks including future research topics will be given in Section 5.

2 Conceptual framework

2.1 Central concepts

Identity disclosure, or re-identification, occurs when a hypothetical *intruder* can determine the identity of the *subject* (individual or other unit) of a record or a data item. *Attribute disclosure* occurs when the intruder can determine (or estimate) the value of a sensitive attribute for an identified subject. Note that the term *inferential* or *prediction* disclosure is sometimes used when disclosure is associated with some non-negligible uncertainty.

Following Skinner (2009), the *confidentiality* of a subject might be said to be protected, if the risk of identity and attribute disclosure is sufficiently low for this subject. On the one hand, it is not a sensible aim to completely eliminate the disclosure risk, if the data are to have any use at all. On the other hand, some method of SDC is needed, in order to keep the disclosure risk sufficiently low. There are broadly two approaches, referred to as *safe setting* and *safe data* (Marsh et al., 1994). The safe setting at the NSO regarding data collection and processing is a different matter to that for dissemination of statistical outputs. In this paper, we focus on safe data during collection and processing.

There are then at least four *parties* relevant to the discussion: one or several *data suppliers*, the *subjects* of the data, the *NSO* and a hypothetical *intruder*. The party of data

suppliers, who acts on behalf of the data subjects, exists rarely in SDC for survey data, whereas the suppliers of administrative data tend to have a different status and interests than many suppliers of transaction or remote sensing data who operate commercially.

The concepts described so far are well-established in the literature of SDC. Below we describe some other important concepts to our approach, which arise from the field of privacy-preserving computation. According to PPT Task Team (2019), set in the context of multi-party computation, *input privacy* means that a Computing Party (i.e. the NSO) cannot access or derive any input value provided by Input Parties (i.e. data suppliers), nor access intermediate values or statistical results during the processing of data "unless the value has been specifically selected for disclosure". In addition, *policy enforcement* is implemented if the Input Parties can exercise positive control on which computations can be performed by the Computing Parties on sensitive inputs, and which results can be published to which Result Parties (e.g. users of official statistics or researchers).

Clearly, input privacy as such would be unacceptably restrictive in the context of secure data collection and processing for official statistics. It is also clearly unacceptable for data suppliers to have control when it comes to the outputs of official statistics. Meanwhile, since it is reasonable to avoid any undue damage to the commercial interests of the data suppliers, we propose to incorporate an element of policy enforcement in the protocol (Section 2.2), whereby a mutual agreement is reached between the NSO and each data supplier and the means of confidentiality protection established.

A couple of additional notes on the terminology are necessary. First, the term "input" here needs to be distinguished from that in "input SDC" (e.g. Elliot and Domingo-Ferrer, 2018), where the disseminated statistical output (by the NSO) can be viewed as input to the receivers. Next, privacy and privacy-preserving are widely used terms in the fields of multi-party computation and data linkage. There are however reasons for distinguishing between confidentiality (which concerns data) and privacy (which concerns data subjects). For instance, as Elliot and Domingo-Ferrer (2018) point out, one should not assume that one has adequately protected privacy by controlling disclosure.

Finally, the concept of TEE in privacy-preserving technology is central to our approach. TEE is conceived as a processing environment that runs *alongside* the standard operating system, such as Android. It provides isolated execution of applications while securing the relevant data and cryptographic keys. In comparison, a debit card with an integrated circuit chip is a familiar example of a different technology, called *secure element*, which has more limited functionality compared to TEE. GlobalPlatform (2011) specifies a TEE standard, although currently there still exist several definitions of TEE, as well as a range of TEE implementations that are only partially compliant with this standard or even non-compliant (e.g. Intel SGX). Nevertheless, Sabt et al. (2015) point to the separation kernel as a fundamental element that is shared across this family of environments, to which we return in Section 3.1.

2.2 Protocol

We shall use the term *protocol* to refer to the agreed formal procedure or system of rules governing secure data collection and processing, which allows for implementing policy

enforcement based on mutual agreements between the NSO and the data suppliers.

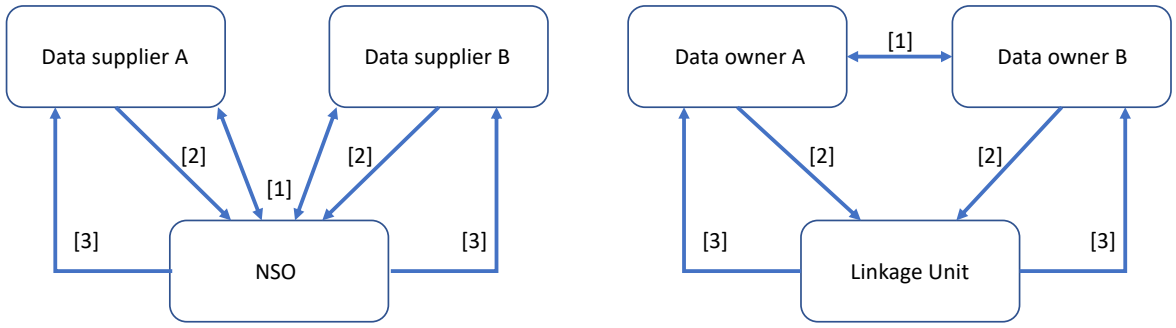

Figure 1: Protocols for secure big data collection and processing (left) and three-party linkage of sensitive data (right). Communication steps: [1] agreement, [2] specified data packages, [3] feedbacks (left) or linkage results (right).

Our proposed general protocol for secure big data collection and processing is shown in the left part of Figure 1. For comparison, the well-established three-party protocol for linking sensitive data is shown to the right, which is reproduced from Figure 4.1(a) of Christen et al. (2020). As we explain below, the two protocols are very different to each other although both ostensibly contain three communication steps.

Under the protocol for three-party linkage, two data owners perform linkage of their sensitive data using a third party, known as the *linkage unit*. At the first step, the two data owners make *agreement* between themselves, including the preprocessing, encoding and encryption methods and the associated secret keys. At the second step, only the specified linkage keys are sent as *data packages* from the data owners to the linkage unit, but not any attribute data otherwise. At the third step, having performed the linkage as agreed, the linkage unit sends the record identifiers for the pairs classified as matches back to the data owners as *linkage results*. The data owners can now either exchange the matched records based on these record identifiers, which have nothing to do with the identifiers of the data subjects, or send them to a data consumer.

For the proposed protocol for secure big data collection and processing, there can be *any* number of data suppliers depending on the situation. The illustration with two data suppliers in Figure 1 is used to underline that generally linkage across different sources may be needed. At the first step, mutual agreement is now made *between the NSO and each data supplier*, which enables a data supplier to exercise policy enforcement pertaining to data collection and processing. It is also possible to restrict certain statistical outputs that are in conflict with the commercial interest of the data supplier, as long as it is based on mutual agreement. At the second step, each data supplier sends the specified data packages — to be described in detail in Section 3. At the third step, feedback is sent from the NSO to a data supplier, provided that it is specified in the agreement between them. For example, the feedback could be a report on the input data quality or an overview of the processed results with respect to the protocol. It may also be possible to send back certain statistical outputs, which make use of the input data and are of interest to the data supplier; these would need to satisfy the confidentiality control applied to any outputs disseminated by the NSO.

We notice that the proposed protocol is structured in the same way as the traditional protocols for survey and administrative data. For instance, formally there exists a mutual agreement between the NSO and each survey unit, by which consent is obtained from the data subject and a pledge of confidentiality is made by the NSO. However, the contents of communication would differ across the protocols. For instance, policy enforcement by agreement is not a traditional concept; or, the NSO sometimes gives a monetary reward as feedback to a survey respondent, but such an incentive is generally deemed inappropriate for administrative or big data suppliers. Above all, data delivery at the second step will be very different, as to be described in Section 3.

We notice also that successful execution of the proposed protocol requires efforts from both the data suppliers and the NSO. In practice, it would be helpful to strengthen the routines or mechanisms, which can ensure both smooth data transfer (e.g. when iteration is necessary due to accidents) and the fidelity of implementation to the agreed protocol. Detailed elaboration of these elements are beyond the scope of this paper.

2.3 SDC as an optimisation problem

Following Skinner (2009), let D be the survey data and $f(D)$ the statistical output resulting from SDC method $f(\cdot)$. Let $R[f(D)]$ and $U[f(D)]$ be, respectively, the measures of disclosure risk and utility of $f(D)$. The data D has both maximum utility and maximum disclosure risk, such that for any candidate SDC method one can assume generally

$$R[f(D)] \leq R[D] \quad \text{and} \quad U[f(D)] \leq U[D].$$

This allows Skinner (2009) to formulate the trade-off between disclosure risk and utility of disseminated statistical outputs as the following optimisation problem.

$$\begin{gathered} \text{For given } D \text{ and } \epsilon, \text{ find } f(\cdot) \text{ which maximises } U[f(D)], \\ \text{subject to } R[f(D)] < \epsilon. \end{gathered} \tag{1}$$

As Skinner (2009) notes, even when it is difficult to specify $R(\cdot)$, $U(\cdot)$ and ϵ unequivocally in a given situation, the optimisation problem (1) can still serve as a conceptual motivation when comparing different SDC methods.

Table 2: Type of SDC method for statistical outputs by format.

Format	Type of SDC method
Table	Non-perturbative (cell suppression, variable recoding), Perturbative
Query	Query restriction, Query perturbation
Microdata	Masking (non-perturbative, perturbative), Synthetic (i.e. artificial) data

Elliot and Domingo-Ferrer (2018) classify existing SDC methods given the format of the statistical output, as summarised in Table 2 here, where a perturbative method operates by introducing noise or distortion in the output, whereas the output would be

coarsened or less detailed although it can remain truthful by a non-perturbative method. Generally speaking, in terms of the optimisation problem (1), any good SDC method should aim to produce the output $f(D)$ whose utility is as close as possible to that of D, while providing sufficiently strong confidentiality protection when the output is sensitive.

Although perturbative SDC methods are useful for disseminating statistical outputs, they are inadmissible for secure big data collection and processing, if they would distort the resulting official statistics. For example, the processed population census data should not distort the ethnicity statistics through intentional perturbation, although perturbation may be applied to a released census table or public query of the census results on the topic of ethnicity. Meanwhile, certain non-perturbative methods may be acceptable, such as restrictions against any query for a single payment transaction, when such queries are unnecessary because the purpose is, say, to make retail turnover statistics. To be able to assess SDC methods generally for data collection and processing, we propose to formulate the trade-off between disclosure risk and utility of big data (to the NSO in producing official statistics) as the following optimisation problem.

Denote by D_{input} the data to be collected and processed, and denote by $g(D_{input})$ an SDC method that is applicable during data collection and processing.

$$\begin{gathered}\text{For given } D_{input} \text{ and } u, \text{ find } g(\cdot) \text{ which minimises } R[D], \\ \text{subject to } U[D] \geq u, \text{ where } D = g(D_{input}).\end{gathered} \tag{2}$$

As with SDC for outputs, even when it is difficult to specify $R(\cdot)$, $U(\cdot)$ and u unequivocally in a given situation, the formulation (2) can still be used to compare different SDC methods $g(\cdot)$. Note that the output $D = g(D_{input})$ resulting from (2) may be the basis of disseminated outputs and therefore D may be subject to (1) later on. In other words, the two optimisation problems (1) and (2) serve different purposes generally.

3 A compartmented system

3.1 TEE mechanism

As mentioned in Section 1, the NSO relies routinely on pseudonymisation for confidentiality protection during collection and processing of survey and administrative register data. Figure 2 illustrates the typical data flows from *input data* to *statistical data*, separated by horizontal dotted lines. As indicated in Figure 2, the input data may be combined with other statistical data at the NSO, to which editing, imputation, weighting and other statistical methods can be applied for *Processing*. Despite pseudonymisation, disclosure risks exist in two particular respects.

r1. Input survey data and most input administrative register data are organised around statistical units, for which a unit ID (or pseudo-ID) is used for linkage and, possibly, process administration, including response chasing or recontact in surveys.

r2. It is customary that all the attributes of the same unit stay together during data processing, and they are possibly supplemented by combining with other data sources.

Thus, pseudonymisation is clearly not a solution to the SDC optimisation problem (2).

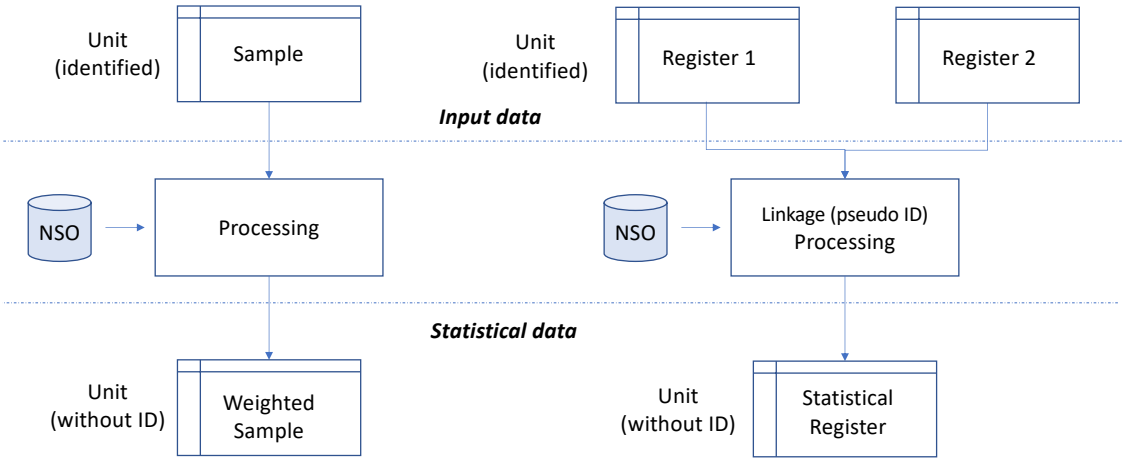

Figure 2: Flows of survey and register data during collection and processing.

We propose to implement the TEE mechanism as defined below, which only makes use of non-perturbative SDC methods, in order to address the optimisation problem (2). The implementation pertains to data delivery and the subsequent necessary tasks before feedback in the governing protocol (Figure 1).

First, the "separation principle" (Kelman et al. 2002) is a mechanism for linking sensitive data, where each participant (or involved party) has access only to the data that are necessary to its role defined in the protocol. In the three-party linkage protocol (Figure 1, right), the linkage unit has access only to the agreed linkage keys but not any other attributes, the data owners have access only to the record identifiers of the matched pairs but not the data subject identifiers of the other party, and the final user has access only to the linked attributes but not any identifiers. The separation principle can be considered to operationalise the *data minimisation principle* of the GDPR.

Next, as a technique for secure computation, TEE implements the "separation kernel" first introduced by Rushby (1981). Basically, the system is divided into several *partitions*, with a strong isolation between them, except for a carefully controlled interface for communications between the partitions. The idea is clearly similar to the separation principle. In addition, the process owner is denied access to the data within the most security-critical processes, which can be considered as an extension to the separation principle.

We shall refer to data isolation according to the separation principle for data linkage (between identifiers and attributes, as well as between different attributes) *and* process (hence, data) access restriction in TEE technology as the *TEE mechanism*.

3.2 System outline

We have developed a compartmented system as a general approach to implement the TEE mechanism for secure data collection and processing. As shown in Figure 3, the system consists of four *chambers* (or partitions): *Extract, Linkage, Grouping* and *Result*. Any input data is divided into separate packages of unit IDs (where available) and attributes, before they are delivered to the NSO. Only the results of each chamber are transferred

between them, while the data in separate chambers and processes are inaccessible to each other. Moreover, statisticians at the NSO are denied access to all the chambers (with solid outlines in Figure 3) except Result, so that it is not possible for them to obtain the microdata that are being processed in the chambers Extract, Linkage and Grouping. Thus, the approach deals with the SDC optimisation problem (2) by removing the aforementioned critical aspects (r1) and (r2) related to the data flows under pseudonymisation (Figure 2), such that disclosure risks are greatly reduced.

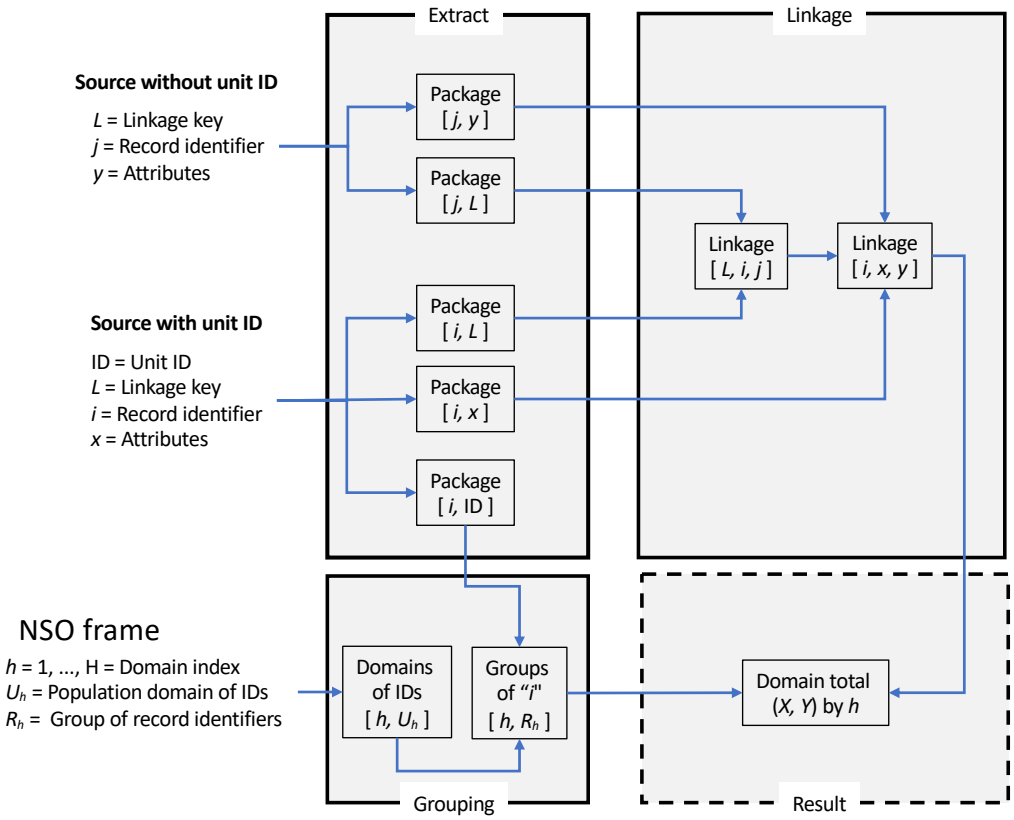

Figure 3: Compartmented system for secure data collection and processing.

As a motivating example, consider combining (debit card) payment transactions from supplier A, which is a source with unit ID (Figure 3) in this case, and (receipts of) purchase transactions from supplier B, which is a source without unit ID. From each payment transaction one obtains the total monetary value of the purchase, denoted by x and the cardholder ID, and given each purchase transaction one can break x down by relevant products (price, quantity and value), denoted by y. Let D_{input} contain the data $[\text{ID}, x]$ from supplier A and $[x, y]$ from supplier B. Let D^* be the joint data $[\text{ID}, x, y]$ per payment-purchase transaction, which is sensitive. Let $g(D_{input})$ be the results, denoted by $[h, X, Y]$, which target the expenditure distribution (over the product groups) for different population domains (i.e. subpopulations), say, age-groups $h = 1, .., H$. We shall naturally assume that $[h, X, Y]$ is not sensitive.

For the SDC optimisation problem (2), we require $g(D_{input})$ to be the same as what one can obtain from D^* directly without applying any SDC method, given which $U[g(D_{input})]$ achieves the required utility. Below we explain how this can be accomplished using the

compartmented system. The corresponding data flow is shown in Figure 4, which allows for a more direct comparison to the approach of pseudonymisation.

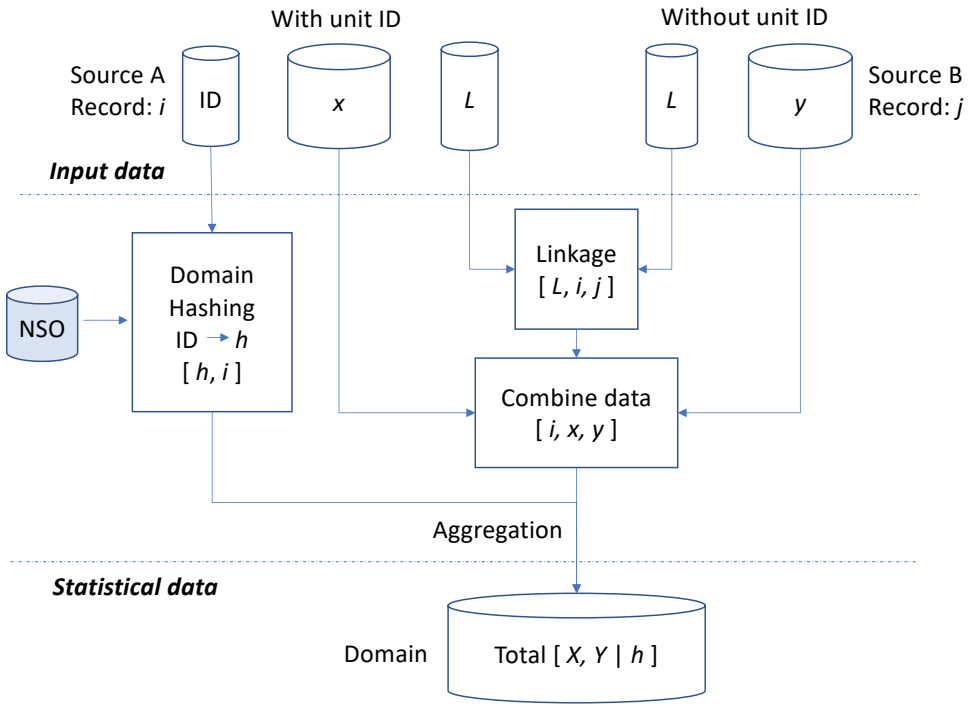

Figure 4: Flows of data through compartmented system.

Extract The input data are *not* data matrices organised around identified units.

Here, the payment transactions from supplier A are a source with unit (cardholder) ID, which allows for disaggregation of expenditures by population domains h. As can be seen in Figure 3, unit ID and attribute x are sent in separate data packages, organised around the payment transactions, with a common but *ad hoc* record identifier i that has nothing to do with the unit ID. It is the record identifier that will follow the data from a supplier. For linkage to the purchases from supplier B, a de-identified key (i.e without any direct unit identifier) is created, denoted by L, which is composed of <time, outlet, value> of each payment-purchase transaction. Note that such de-identified linkage keys are often possible for linking big data sources, as will be discussed below under Linkage.

Similarly, purchase transactions from supplier B are separated into data packages of linkage key L and attribute y, where the record identifier j is unrelated to i above.

Notice that the products and the associated attributes (price, quantity, value) are not sensitive on their own, as long as they are separated from the cardholder ID. In situations with rich attributes, such as when one of the sources to be linked is the population census, one can divide all the attributes into separate input data packages, say, y_1, y_2, ..., y_K, each of which is associated with the same record identifier, in order to reduce the disclosure risks. The following operations remain otherwise the same.

Linkage Only the attributes that need to be processed jointly are to be linked, and linkage is virtually based on the three-party protocol governed by the separation principle.

To start with, de-identified linkage keys are possible for linking big data sources, when the underlying transactions (or other events) are not directly based on unit ID in reality. Linking debit card payment and purchase transactions above provides an example. As another example, consider invoice-based business-to-business (B2B) payments, say, in two different sources: source A of the invoices and source B of the actual payment transactions. The matched data would contain attributes such as due date, objective, quantity, value/VAT, etc. They would involve identified units of businesses and bank account owners. However, to link the invoices from supplier A and the payments from supplier B, one can just use the invoice number as the de-identified linkage key.

Of course, in some situations unit ID may still need to be the linkage key, such as when linking the total of person-to-business payments of each business from a data supplier to the sample of a related business survey conducted by the NSO. The NSO can implement the three-party linkage protocol virtually in the Linkage chamber, where the pairing of matched encrypted unit IDs is separated from the joining of the attributes associated with the linked records. Thus, as can be seen in Figure 3, whether or not the linkage key is de-identified, the linkage operation is completed in two steps:

I. linking only the record identifiers across the two sources: $[L, i, j]$,

II. joining the attributes to yield $[i, x, y]$ via the matched record identifiers above.

Note that only the record identifier i from source A (with unit ID) is retained in the linked data, because only the unit ID is needed for generating results by population domains. The record identifier j from source B is dropped to minimise the number of values associated with the linked objects. Next, step I above only needs to be performed once, whereas step II can be repeated to create multiple linked datasets of different attributes. Finally, the Linkage chamber is implemented as a TEE-like enclave, so that the NSO statistician only has access to the linked data $[i, x, y]$ afterwards, but not the input data packages or the intermediate data set $[L, i, j]$.

Grouping Domain classification of unit ID is always separated from linkage of attributes.

The domains $h = 1, ..., H$ are defined based on the NSO's relevant population frame. Each U_h is the set of unit IDs belonging to the corresponding population domain. The transformation from unit ID to domain index is formally a hash function from $[\text{ID}, i]$ to $[h, i]$, which yields each R_h as the set of the corresponding record identifiers i. The Grouping chamber is also implemented as a TEE-like enclave.

In cases where de-identified linkage keys can be used for linking data across sources, the Grouping chamber is the only place in the system, which uses the input unit ID. Interception of $[h, R_h]$ outside the Grouping chamber would not lead to identification of the in-source units (e.g. cardholders here), since R_h contains only the record identifiers.

Result We have now $[h, i]$ from the Grouping chamber and $[i, x, y]$ from the Linkage chamber. This allows one to produce the corresponding domain cross-classified counts (X, Y), i.e. the expenditure distribution for each specific age-group here.

Note that, in this example, the resulting expenditure distribution $g(D_{input}) = [h, X, Y]$ is biased due to the unavoidable errors of coverage and classification (of product groups), but has virtually no variance compared to estimating the target distribution based on the traditional Consumer Expenditure Survey. Zhang (2020) develops audit sampling inference for such big data statistics.

4 Opportunities and illustrations

Below we give some illustrations of the various ways of configuring the compartmented system and the resulting $g(D_{input})$ in different formats.

4.1 Sample contingency tables

The expenditure table $g(D_{input})$ in Section 3.2 is based on *all* the data in the relevant sources. Below is an example of $g(D_{input})$ as sample contingency tables.

Register-based employment statistics in recent years clearly show a large increase in the number of people with multiple jobs in Norway. Suppose one is interested in the statistics of job-related travel patterns of people with multiple jobs. On the one hand, correct classification is not always possible based on Register data. For instance, an employer may have multiple locations, all of which are not necessarily included in the Business Register. But even when they are, it is not always clear where an employee works, if the work place locality of an employee is not directly recorded in the administrative sources. On the other hand, mobile phones can provide accurate location and movement data. But it is not always clear whether the presence at a given location is job-related, or if the algorithm can correctly handle the temporal variations of travel patterns.

Schenkel and Zhang (2020) develop a method for adjusting the misclassification errors of two fallible classifiers observed in a non-probability sample, for which one only needs contingency tables of the joint classification, but not linked data at the individual level. The compartmented system (Figure 3) can be used to obtain sample two-way tables (X, Y) of the two classifiers, where x is based on the Register data at the NSO and y on mobile phone locations. One only needs to configure the Extract chamber as follows.

- The NSO draws a sample from the target population given the statistical reference time point of interest, which is the input source with unit ID. Send only the list of telephone numbers associated with the sample to a telephone company.

- The telephone company is only able to classify a subset of the telephone numbers it receives, not least because it does not have a monopoly in the market. The company returns its classifications pertaining to the reference time point and the associated telephone numbers, in separate data packages.

The linkage key is the encrypted telephone number. The resulting $g(D_{input})$ are domain-specific two-way tables (X, Y). Disclosure risks are low both ways: the NSO does not have access to the individual classifications by the telephone company, nor the other way around. Moreover, the telephone company cannot validly estimate its market share based

on the telephone numbers that belong to its customers in the NSO sample, since it does not know the sampling design. If deemed appropriate and agreed in the protocol, the NSO can send the estimated misclassification errors (of y based on mobile phone locations) to the telephone company, as a feedback (Figure 1). We refer to Schenkel and Zhang (2020) for the details of the estimation method.

4.2 Big data as auxiliary information for sample surveys

To fix the idea, consider retail turnover statistics. Denote by s the sample of business units taken by the NSO, where π_k is the sample inclusion probability of $k \in s$. To simplify the elaboration, suppose unstratified sampling design here; the adaption to stratified sampling is straightforward except some extra complications of the notation. Let y_k be the turnover excluding VAT for $k \in s$, which is the target measurement. Let x_k be the total of debit card payments to $k \in U$, where U is the population of business units. The ratio estimator of the total turnover $Y = \sum_{k \in U} y_k$ is given by

$$\hat{Y} = X \frac{\sum_{k \in s} y_k/\pi_k}{\sum_{k \in s} x_k/\pi_k} \ .$$

Note that x_k is the sum of all debit card payment transactions, denoted by Ω_k for $k \in U$, where x_{ki} is the payment associated with transaction $i \in \Omega_k$, such that $x_k = \sum_{i \in \Omega_k} x_{ki}$. Due to the different VAT applied to different products, x_k would have been neither equal nor proportional to y_k, even if all the payments had been by debit card.

Let $g(D_{input})$ consist of pseudonymised sample data $\{(y_k, x_k) : k \in s\}$ and the auxiliary total X, given which the utility is considered to be achieved for the SDC optimisation problem (2). The auxiliary big data over all the payment transactions are given by

$$D_B = \{x_{ki} : k \in U, i \in \Omega_k\}.$$

Suppose it is agreed that disclosure risk is sufficiently low if the NSO obtains $g(D_{input})$ but not D_B. The compartmented system (Figure 3) can be configured as follows.

- *Extract* Let the debit card payment transactions be the source with (business) unit ID and attribute x_{ki}. Let the NSO sample be the second source with attribute y_k. The linkage key is the encrypted unit ID.

- *Linkage* Record linkage here amounts to linkage of the records from both sources first, and de-duplication/aggregation of x_{ki} to x_k by the linkage key afterwards. The linked data are $\{(x_k, y_k^*) : k \in U\}$, where $y_k^* = y_k$ if $k \in s$ and y_k^* is missing if $k \notin s$.

- *Grouping* Each R_h consists of the encrypted unit IDs, for domain $h = 1, ..., H$.

- *Result* Aggregation of x_k over the linked data yields X. Dropping the linked data with missing y_k^* yields the pseudonymised sample data $\{(y_k, x_k) : k \in s\}$.

4.3 De-identified housing rental microdata

The (housing) Rental Market Survey has two well-known practical difficulties. First, the population of housing rental objects is typically unknown in its entirety based on the available Register data on population, address, tax, etc. Second, surveying and tracking a sample of rental objects can be resource demanding, especially at the time an address is first drawn into the sample and following a change of tenants. Since the rents are mostly paid by standing orders in online banking, combination with such Transaction data can potentially provide a more efficient and cost-effective approach, as discussed below.

To fix the idea, let N_A be the number of potential housing rental objects in the NSO's sampling frame based on Register data. Let the true number of rental objects be given as $N = pN_A$, where p is an unknown constant. Let the size of the initial address sample be n_A. Again, for simplicity of elaboration, we do not explicate any stratified design here, although an adaption is straightforward in concept. Suppose that n of the addresses are verified to be rental objects after canvassing, where $E(n) = n_A p = n_A N/N_A$ with respect to the sampling design. An estimator of N is then given by

$$\hat{N} = N_A n/n_A$$

whose sampling variance is given by

$$V(\hat{N}) = \frac{N_A}{n_A}(N_A - n_A)p(1-p)$$

Next, let M be the number rental objects among the N_A frame units, for which standing orders for rent can be found in the payment transactions data. This requires connecting the payer and the address of each standing order in the source, which is sensitive. The compartmented system can be used for confidentiality protection, as will be explained further below. For the moment, let us illustrate first how in principle the data can be combined with the sample above administered by the NSO.

Suppose m among the M objects can be matched to the n sample rental objects, where the linkage key is address. One can now greatly reduce the cost of surveying the tenants or owners for these matched objects, both when the address is first drawn into the sample and following a change of tenants. Moreover, let $\theta = M/N$ be the unknown constant, such that $E(m) = \theta p n_A$ with respect to the sampling design. We have

$$\frac{E(n)}{E(m)}M = \frac{pn_A}{\theta p n_A}\theta N = N$$

such that the Lincoln-Petersen estimator of N is given by

$$\tilde{N} = nM/m$$

Note that the estimator $\tilde{N}$ is traditionally motivated from a model-based perspective, where (n, M, m) are all considered as random variables; see e.g. Wolter (1986). Zhang (2019) establishes model consistency of $\tilde{N}$ by treating one of the sources, say, M as fixed,

so that only (n, m) are random variables. This conditional approach to inference can be applied to the setting here, where the randomness of (n, m) derives only from the sampling design. An approximate sampling variance of $\tilde{N}$ is given by

$$V(\tilde{N}) \doteq \frac{M}{E(m)^3} E(n)E(n-m)\big(M - E(m)\big) = \frac{N_A}{n_A}(N_A - n_A)\frac{p}{\theta}(1-\theta)$$

It follows that the relative efficiency (RE) of $\tilde{N}$ against $\hat{N}$ is given by

$$\mathrm{RE} = \frac{V(\tilde{N})}{V(\hat{N})} = \frac{1-\theta}{(1-p)\theta}$$

In the extreme case of $p = 1$ while $\theta < 1$, where the NSO's sampling frame is perfect, the estimator $\hat{N}$ would naturally have zero variance and dominates the estimator $\tilde{N}$. At the other end, in the extreme case of $\theta = 1$ while $p < 1$, where one can identify the objects perfectly using the Transaction data, the estimator $\tilde{N}$ would naturally have zero variance and dominates the estimator $\hat{N}$. As an illustration of a practical situation, according to the information available at ssb.no, the initial sample of the Rental Market Survey 2020 has $n_A = 35286$ and $n = 9727$, giving $\hat{p} = 0.276$. It follows that the estimator $\tilde{N}$ using Transaction data could be more efficient than the one-sample estimator $\hat{N}$ if

$$\theta > 1/(2 - \hat{p}) = 0.580$$

i.e. roughly at least 60% of the rental objects can be identified from the standing orders.

Let the input data D_{input} consist of two microdata sets. The first one is the NSO's sampling frame, where each record i corresponds to one of the N_A addresses, with associated dwelling physical characteristics x_i from the Register data at the NSO, including the address, the locality (such as postcode), whether the address is known to be a housing rental object from the past, and the sample indicator $\delta_i \in \{0, 1\}$. The second one arises from Transaction data, where each record j corresponds to one of the M rental objects, the address and the associated rent y_j (possibly over several months) according to the algorithm for classifying the standing orders.

Let $g(D_{input})$ be the set of microdata from linking the two input data sets, where the linkage key is the address. Each record in $g(D_{input})$ is given by $(i, x_i, \delta_i, y_i^*)$, where i is the record number in the NSO's sampling frame, and y_i^* is either the rent identified in the Transaction data or missing, denoted by $y_i^* = 0$ for convenience. Since $g(D_{input})$ contains neither person ID nor address, we consider it to have sufficiently low disclosure risks. It has the same utility for Rental Market statistics as the microdata set containing additionally both person ID and address. For instance, we have

$$n = \sum_{i=1}^{N_A} \delta_i \quad \text{and} \quad M = \sum_{i=1}^{N_A} \mathbb{I}(y_i^* > 0) \quad \text{and} \quad m = \sum_{i=1}^{N_A} \delta_i \, \mathbb{I}(y_i^* > 0)$$

While it is possible for the NSO to carry out additional operations on the sample addresses with missing rents in the Transaction data based on the corresponding record numbers i,

the interception of i in $g(D_{input})$ does not carry any high risk of identity disclosure.

The compartmented system (Figure 3) can be configured as below to obtain $g(D_{input})$.

- *Extract* The attribute data package derived from the NSO's sampling frame contains (i, x_i, δ_i) with N_A records, whereas that from the Transaction data supplier contains records (j, y_j). Each source sends a data package of linkage keys (i, L_i) and (j, L_j). Unit ID is not otherwise needed in this case.

- *Linkage* The linkage key is the address. The Transaction data records that cannot be matched are removed. The linked records $(i, x_i, \delta_i, y_i^*)$ are as defined above.

- *Grouping* No domain classification is needed, beyond what can be determined given the preprocessed locality and dwelling physical characteristics in x_i.

- *Result* Microdata set $g(D_{input})$ as defined above.

Record linkage here has removed any addresses in the Transaction data which are outside the NSO's sampling frame. One can apply the usual editing procedures for outliers or values that are otherwise deemed inappropriate. The number of objects with rents from the Transaction data would be much larger than the sample size n of any Rental Market Survey. The objects with changing tenants are automatically tracked over time by this approach. Finally, the NSO retains the choice whether and what to do with the $N_A - M$ addresses for which no rents are identified in the Transaction data.

4.4 Network business structure

The trade (or other) relationships among the businesses can be represented by a graph, denoted by $G = (U, A)$, where the businesses are the nodes in U, and a directed edge from node k to node l exists if business k sells to l, denoted by $(kl) \in A$. Each connected sub-graph of G can be referred to as a network. Values can be added to G to form a valued graph, e.g. x_{kl} can be the total value of sales from k to l over a given period and $x_k = \sum_{l \in U} x_{kl}$ the total turnover of business k in the economy defined by U. One can e.g. let $z_k = 1$ if business k trades internationally and 0 otherwise.

Various network parameters can be used to describe the business structure represented by G. For instance, density and transitivity are at the maximum value 1, if the structure of the network is complete, where everyone sells to everyone else. Or, a measure of the openness of the economy can be defined based on the lengths of the short paths that connect each business to one that is engaged in import or export (with $z = 1$). See e.g. an analysis of the Belgian production network (Dhyne et al., 2015), where the data are based on mandatory reporting to the Belgian tax authority.

Let $g(D_{input})$ be the pseudonymised valued graph G, where the values $\{x_{kl} : k \neq l \in U\}$ are based on the business to business (B2B) invoices, given which the utility is considered to be achieved for the SDC optimisation problem (2). The business unit ID associated with the nodes in $g(D_{input})$ is the master key of the NSO, which is meaningless to any outsider.

Let Ω_{kl} be all the invoices from seller k to buyer l, such that $x_{kl} = \sum_{i \in \Omega_{kl}} x_i$, where x_i is the value of invoice i. Let the invoice data D_B be the valued multigraph G_B, where Ω_{kl} are the edges from k to l in G_B, with associated values $\{x_i : i \in \Omega_{kl}\}$. Suppose it is agreed that disclosure risk is sufficiently low if the NSO obtains $g(D_{input})$ but not D_B.

Let there be two data suppliers of B2B invoices and payment transactions, respectively. The linkage key L between them is the invoice number, the purpose of which is to remove the void invoices due to outstanding payment. Each invoice i is associated with the business ID of both the seller k and the buyer l, the invoice value x_i, the invoice data and due date. Each payment j is associated with the payment value and date.

The compartmented system (Figure 3) can be configured as follows.

- *Extract* One can create three separate data packages for the invoices: (a) the seller ID k and record number κ, (b) the buyer ID l and record number ℓ, and (c) the invoice i with the invoice value and dates, the associated record numbers (κ, ℓ) and the linkage key L. Only one data package of is needed for the payments, with the linkage key L and payment value and date.

- *Linkage* Record linkage in this case amounts to de-duplication of the data package (c), whereby all the non-void invoices (i, x_i) are accumulated for each distinct pair of (κ, ℓ). The results can be given in three data tables: the valued adjacency matrix $[x_{\kappa\ell}]$, the sellers κ with the associated sales value x_κ, and the buyers ℓ with the associated expenditure value x_ℓ. The data packages (a) and (b) are simply retained.

- *Grouping* The data package (a) of seller ID can be used to group the record number κ into population domains, say, by NACE and export/import status. Similarly for domain classification of the buyers using the data package (b).

- *Result* The pseudonymised valued graph $g(D_{input})$ can be created, where each node is associated with the seller record number κ, or the buyer record number ℓ, or both.

Various networks can be constructed given $g(D_{input})$. For instance, for a sales network of the businesses in a given NACE group U_h, one can start with all the edges that originate from the nodes in U_h, and include wave-by-wave the down-stream edges (and businesses) that originate from the additional nodes in the previous wave, until no more edges can be added. Transitivity of sales relationships or network 'distance' to export can now be analysed for the businesses in U_h based on this network.

Unlike a mandatory reporting system of such detailed information, using Transaction data does not place any burden on the business community, and the subsequent analysis of business structure is not based on identified units.

5 Final remarks

We have formulated SDC for data collection and processing as an optimisation problem, for which pseudonymisation is not a solution. We have also developed a flexible approach based on the compartmented system that implements the TEE mechanism, together with

the general protocol governing the communications between the NSO and the data suppliers. As we have illustrated, the system can be configured to obtain various table data and microdata sets originated from non-survey big data sources.

With respect to the three benign parties involved, the proposed approach provides much stronger protection of the confidentiality of the data subjects during data collection and processing. It provides a mechanism for avoiding any undue damage to the commercial interests of the data suppliers, or their pledges of confidentiality protection to their users or clients who are the data subjects. In return we argue that this would make it easier for the NSO to gain trust and acceptance from data owners, stakeholders and the public, and thereby smooth the access to useful new big data sources.

The hypothetical malicious intruder would basically need to gain control over the entire compartmented system, in order to achieve what otherwise could have been achieved just by gaining access to the pseudonymised dataset during collection and processing. For instance, intercepting any dataset between Linkage (or Grouping) and Result chambers does not give the intruder access to any identifiers, whereas intercepting the dataset between Extract and Grouping chambers would only give the (pseudonymised) identifiers but without any associated attributes.

The formulation (2) can provide a rigorous basis for comparing different configurations of the compartmented system, given suitable measures of the associated disclosure risk. Various measures of disclosure risks have been proposed in the context of SDC for dissemination of statistical outputs. See e.g. Fuller (1993), Skinner and Elliot (2002), Skinner and Shlomo (2008). However, their relevance to the compartmented system is not obvious. For instance, the TEE mechanism implements access restriction to microdata in the Linkage chamber, given which the disclosure risk of any unique data record would seem to have been removed by stipulation. Developing formal measures of disclosure risk $R(\cdot)$ that are relevant to (2) is a topic for future research.

Our focus in this paper has been on confidentiality protection during big data collection and processing. We have also indicated that statistical methods for adjustments and uncertainty assessment are generally required. More broadly, as discussed by Zhang et al. (2020), a greater emphasis should be given to *statistical design*, which deals conceptually with two central questions: (i) which data to collect, (ii) how to collect and use the data. It is essential to have a clear idea *before* contracting the protocol with a big data supplier, not only about the potential uses of a given source throughout the statistical system, but also the possibilities of combining it with other sources. Systematic statistical designs are important to develop a sustainable official statistical system, where non-survey big data sources are becoming more and more important.

References

[1] Christen, P., Ranbaduge, T. and Schnell, R. (2020). *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer International Publishing.

[2] Dhyne, E., Magerman, G. and R. Stele (2015). *The Belgian production network 2002-2012*. National Bank of Belgium, Brussels. http://aei.pitt.edu/97432/1/wp288en.pdf

[3] Elliot, M. J. and Domingo-Ferrer, J. (2018). *The Future of Statistical Disclosure Control*. Paper published as part of The National Statistician's Quality Review. London, December 2018. https://arxiv.org/abs/1812.09204

[4] ESSnet Big Data II (2020). Web portal. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

[5] Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9:383-406.

[6] GlobalPlatform (2011). *TEE system architecture*. http://www.globalplatform.org/specificationsdevice.asp

[7] Goldreich, O. (1998). *Secure Multi-Party Computation*. http://www.wisdom.weizmann.ac.il/~oded/PSX/prot.pdf

[8] C.W. Kelman, C.W., Bass, A.J. and Holman, C.D.J. (2002). Research use of linked health data – a best practice protocol. *Australian and New Zealand Journal of Public Health*, 26:251-255.

[9] Marsh, C., Dale, A. and Skinner, C.J. (1994). Safe data versus safe setting: access to microdata from the British Census. *International Statistical Review*, 62:35-53.

[10] PPT Task Team(2019). *UN Handbook on Privacy-Preserving Computation Techniques*. https://marketplace.officialstatistics.org/privacy-preserving-techniques-handbook

[11] Rushby, J.M. (1981). Design and verification of secure systems. *SIGOPS Oper. Syst. Rev.*, vol. 15, no. 5, pp. 12-21.

[12] Sabt, M., Achemlal, M. and Abdelmadjid Bouabdallah, A. (2015). Trusted Execution Environment: What It is, and What It is Not. *14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. doi:10.1109/Trustcom.2015.357

[13] Schenkel, J.F. and Zhang, L.-C. (2020). Adjusting for two fallible classifiers jointly observed in a nonprobability sample. *Submitted*.

[14] Skinner, C.J. (2009). *Statistical Disclosure Control for Survey Data*. In: Pfeffermann, D and Rao, C.R. eds. Handbook of Statistics Vol. 29A: Sample Surveys: Design, Methods and Applications. pp. 381-396.

[15] Skinner, C.J. and Elliot, M.J. (2002). A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society, Series B*, 64:855-867.

[16] Skinner, C.J. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of American Statistical Association*, 103:989-1001.

[17] Wolter, K. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81: 338-346.

[18] Zhang, L.-C. (2020). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, DOI:10.1111/rssa.12632.

[19] Zhang, L.-C. (2019). A note on dual system population size estimator. *Journal of Official Statistics*, 35:279-283.

[20] Zhang, L.-C., Haraldsen, G., Pekarskaya, T. and Hole, B. (2020). *Non-survey big data for official statistics: Sources, usability and statistical design.* Statistics Norway, Working Documents.

[21] Zuboff, S. (2019). *The Age of Surveillance Capitalism.* Profile Books.