1 Stock take: Overview of existing tools for risk management in view of privacy issues

Table of contents

1 Stock take: Overview of existing tools for risk management in view of privacy issues
   a) Risks to privacy
   b) Data access strategies
      (i) Onsite access
      (ii) Remote access
      (iii) Microdata dissemination
   c) Database privacy technologies
      (i) Privacy-preserving data mining (data owner’s privacy)
      (ii) Private information retrieval (user’s privacy)
      (iii) Remote access statistical disclosure control
   d) Privacy of estimates and microdata
      (i) Microdata
      (ii) Tabular estimates
   e) Evaluation of different privacy approaches
      (i) Utility and disclosure risk for queryable databases
      (ii) Utility and disclosure risk in microdata disclosure control
      (iii) Utility and disclosure risk for tabular data
   f) Privacy software
      (i) Remote server
      (ii) Microdata dissemination
      (iii) Statistical disclosure control software

2 References

1 Stock take: Overview of existing tools for risk management in view of privacy issues

Statistical organizations collect information from people and businesses, here called data providers, in order to support decision-making, research and discussion within governments and the community. This information is typically released to government and the community in the form of estimates or micro-data.

Data providers reasonably expect that disclosure of their sensitive information (e.g. income or health condition) is unlikely. Without this expectation of privacy, data providers may not provide information to the agency. This in turn would undermine the ability of the agency to fulfill its mission. Different people, cultures, and nations have a variety of expectations about what information is sensitive and what constitutes unlikely. In some cases there will be an explicit legal answer to these questions but in many cases, judgement is required.
This chapter reviews current practices by statistical organizations that aim to manage two conflicting aims: to provide access to data for the benefit of society and to meet society’s expectation that sensitive information about data providers will be kept private. This chapter will focus on these practices, and their advantages and disadvantages, in so far as they relate to traditional data sources (e.g. sample surveys and administrative data collected by governments). Section a) broadly discusses how risks to privacy depend on characteristics of the data, how the data are released and the potential motivations to breach a data provider’s privacy. Section b) discusses the different ways in which statistical organizations allow analysts or researchers to access micro-data. Section c) discusses managing the risk of disclosure associated with databases. Section d) is a brief summary of the advantages and disadvantages of the different approaches to managing privacy. Section e) mentions some relevant software, some of which is freely available.

a) Risks to privacy

For simplicity, in this chapter it is assumed that all data are sensitive. The risk of disclosing sensitive information about a data provider is called disclosure risk. For micro-data releases, disclosure could occur by identification or attribute disclosure, even after removing name and address. Identification disclosure can occur when a record is matched to its corresponding record on an external administrative file that contains the data provider’s name and address. This can occur via a quasi-identifier (e.g. in the case of a person this matching key could be age, sex, suburb and occupation) available on both the micro-data and an external administrative file. Attribute disclosure occurs when new information is disclosed about a person or business. Continuing with the above example, following identification of the person, all sensitive variables or attributes contained in the micro-data are disclosed.

For estimate releases, disclosure can occur as for micro-data releases, except that information about a data provider must first be derived from one or more estimates. For example, consider a frequency table with dimensions age, sex, suburb, occupation and income range and that a margin defined in terms of age, sex, suburb and occupation has a count of one. The value of variables that define the margin could be used as a quasi-identifier for the purpose of identification disclosure and the internal cell of the table would reveal the person’s income range, which is attribute disclosure.

The risk of disclosure is also influenced by the type of access. Analysts can directly access micro-data, which means they can observe data values. Direct access could be supervised by official agency staff via onsite access or it could be unsupervised via a micro-data release. In the case of estimate releases, the disclosure risk is affected by whether it is the official agency who decides what estimates are released (e.g. publications) or whether the analyst has some flexibility to decide what estimates are released (e.g. via some form of remote access).

Two broad areas of managing disclosure risk are:

1. Reducing the risk that an attempt at disclosure will be made
2. Releasing data in a manner that is not likely to enable disclosure

Risk 1. refers to the risk that a user is motivated to attempt disclosure, or alternately, that a user is not sufficiently deterred from making an attempt at disclosure. This risk can be managed by masking the data, by selecting which users to grant access to the data, by training users, and by the way in which the data are accessed (e.g. access to micro-data via a remote server). Factors affecting whether a user would attempt disclosure are:

- **Motivation to attempt disclosure.** The motivation could be high if there is the potential for substantial monetary reward from disclosure (e.g. commercially sensitive data) or if the information is highly private or personal (e.g. health conditions, political opinions or personal behavior).

- **Deterrents to attempt disclosure.** These include: requiring users to sign legally enforceable undertakings not to attempt any disclosure, threatening penalties for breaching these undertakings, auditing users to ensure they are using the data for the purpose for which access was granted

- **Effort, skills and technology required to make a disclosure.** With technological advancements, attempts at identification may require little effort or skill. Applying methods of statistical disclosure control will increase the effort and skills required to make an identification. Developing standard ‘profiles’ of users (e.g. private individual, telecommunications company, and government organization) may help identify users for whom the required skills and technology are currently available.

Risk 2 refers to the risk of disclose assuming that an attempt at disclosure was made. The factors affecting this type of risk include:

- **Amount of data accessible (or that can be viewed) by the user.** Is the data available in the form of a unit record file, a set of predetermined aggregates, or in a set of user-defined aggregates (e.g. via a remote server or from user requests);

- **Level of detail in the data.** The data sufficient for identification

- **Presence of remarkable units that are readily identifiable.

- **Accuracy of the collected data:** Reporting errors, poorly defined metadata, data that are masked for the purposes of statistical disclosure control or data that is out-of-date, all reduce the likelihood of identification.

- **Coverage of the data:** A person can only be identified if their information is available in the data. Therefore the greater the chance that the person is in the data, the greater the risk of disclosure for the person

- **Presence of other information that can assist in identification, including:

  - publicly available information;

  - restricted access data holdings that a data user may have access to; and

  - personal knowledge that a user may have.

b) Data access strategies
Historically, national statistical organizations were the only entities collecting, storing, and analyzing large amounts of data for many years. Until the early eighties of the last century direct access for external researchers to the data collected by the organizations was unthinkable. No-one outside the organizations had the capabilities to analyze these data and strict regulations prohibited that anybody not employed at the organizations could have access to the data on the premises of the organization. So the main concern regarding confidentiality was that published tables based on the underlying microdata might lead to the identification and attribute disclosure for individual included in the database. When modern desktop computers became available in the early eighties, researchers at universities and other research institutions suddenly had the capacity to analyze microdata on their own computers and calls for access to data collected by government agencies were raised for the first time. As a response to these calls statistical organizations developed different strategies over the last decades to enable external researchers to analyze their data without violating confidentiality regulations. These strategies can be summarized under three broad topics: microdata dissemination, onsite analysis in research data centers, and remote access.

(i) Onsite access

An alternative approach for providing data access for external researchers emerged at the end of the last century. Many statistical organizations established research data centers (RDC) on their premises. This form of data access is usually not available to the general public. Only academic researchers may use this approach. Generally, external researchers that wish to analyze the data have to submit a research proposal in which they explain their research ideas, discuss which datasets they want to use and why their research questions can only be answered using the specific dataset. Some organizations also require that the researchers determine in advance which variables they will be using. If they realize later that they would need additional variables to answer their research questions a new proposal would need to be submitted. Once the proposal has been submitted, the organizations check whether data access can be granted. In many cases formal approval from other legal entities also needs to be given. For example, at the German Institute for Employment Research (IAB), each requested data access needs to be formally approved by the Ministry of Labour. Once the request has been approved, the researcher signs a contract in which she guarantees that she will not use the data for re-identification purposes. The researcher can then schedule a time slot during which she will visit the organization. She will be provided with computing facilities and access to the requested data. The computing facilities are not connected to the internet and the data and all analysis code can only be stored in selected folders on the desktop machine. All printing and external data storage devices such as USB drives are disabled. The researcher is not allowed to bring any electrical devices such as laptops, cameras, or cell phones inside the RDC and all her activities on the computer are monitored. She will be able to manipulate the data in any way but any output that should be used outside the RDC for example for inclusion in a research article will be carefully scrutinized by staff at the RDC to ensure that no confidentiality guarantees are violated if the output is released.

Compared to microdata dissemination the organization has much better control over what happens with the data. Only accredited researchers will have access, the data never leaves the RDC, the activities of the researchers are monitored and any output taken out of the RDC is carefully checked before release. For this reason, more datasets are usually available on site and often the data are only altered slightly or even not altered at all. An obvious disadvantage of the approach is that it is costly for the researcher and the data providing organization. Researchers have to invest time and money to travel to the RDC and find accommodation for several nights. Often, unanticipated problems arise when working with the microdata and thus the time slot that was requested might not fit. However, a simple extension of the time slot is often not possible since the machines might be booked by other researchers. Thus, the researcher will have to wait until another slot becomes available to finish her research. On the other hand, the approach is cost intensive for the statistical organization since staff are needed to run the facility and to check each output that is to be used outside the RDC.

A recent approach to mitigate some of the costs at least for the researchers are RDC-in-RDC solutions. The idea is that statistical organizations allow researchers to access their data in the safe environment of RDCs run by other agencies. For example, the IAB has set up several RDC-in-RDC solutions throughout Germany and also at several universities in the USA. With this approach the traveling distances for the researchers are greatly reduced and researchers that might not be able to cover the traveling expenses otherwise can now access the data at an RDC that is closer to their own research institution.

(ii) Remote access

The latest development in data access research is remote access. With remote access external researchers connect with a secure server and run their analyses on these machines. The microdata never leaves the secure environment. A predecessor of remote access is remote execution. With remote execution, researchers are provided with some dummy or test datasets. These datasets mimic the structure and content of the original data as closely as possible but no meaningful inferences can be obtained based on the data. Based on these test data researchers develop their analysis code and send the code to the RDC. Staff at the RDC will run the analysis code on the original data. The output is checked manually for confidentiality violations and sent back to the researcher if no violations are detected. Since the analytical validity requirements are much lower than the requirements for microdata dissemination it is much easier for statistical organizations to generate such datasets and disseminate them to the public. Remote execution has already been established at several organizations. However, the output still needs to be checked manually, which means an investment for the RDC. Furthermore, researchers have to wait until the output has been released before they can decide whether further adjustments are necessary. Since the test data will not reflect all subtleties of the original data analysis code might have to be sent back and forth between the RDC and the researcher several times before the final output is obtained.

To avoid these hassles it seems a consequent step to circumvent any manual interference is necessary. However, besides the technical problems such as ensuring a safe connection between the computer of the researcher and the secure server, several other issues complicate the establishment of remote access solutions. The main such problem is that in many countries any output that is shown on the screen of the researcher would be considered a data dissemination. Furthermore, a full remote access solution requires that all outputs that are transmitted to the researcher are guaranteed to confirm with data confidentiality regulations. In many cases it seems trivial to decide whether an output can be classified as safe or not. For example, maximum queries should never be answered since they provide information on individual records. However, it is not always easy to detect problematic queries and misguided data users could use several strategies to hide their attacks on the data (see O’Keefe and Chipperfield (2013) for an excellent review). Furthermore, several studies (see for example Gomatam et al. (2005), Bleninger et al. (2011), or Drechsler et al. (2014)) have shown that seemingly harmless multivariate analyses such as linear regressions could be used to obtain exact information for individuals in the database. Section C (ii) below expands upon this discussion about remote access.

(iii) Microdata dissemination
Ever since the early eighties statistical organizations have been developing strategies to disseminate microdata without violating confidentiality requirements. Since the organizations have no control over who will access the data and what the data are used for once they are disseminated, they need to ensure that the released data are sufficiently protected. In many countries legal regulations require that it is strictly impossible to identify any individual in the released data even if the information that is learned would not be considered sensitive. Disseminating data that still contain valuable information is a challenging task under this requirement. Early protection methods such as swapping (see discussion below) mostly focused on ensuring confidentiality paying little attention on the potentially negative impacts of the approaches on analytical validity but over the years more and more sophisticated methods were developed (see the review below) and nowadays thousands of datasets are available around the world for public use using a variety of methods to ensure confidentiality. A good example is the IPUMS International project (www.ipums.org) from the University of Minnesota that offers access to public use micro data samples (PUMS) of Census data from 74 countries. Still, sensitive microdata such as business data or medical data are often not released. Balancing disclosure risk and analytical validity is especially difficult for these data since re-identification is often easy and at the same time the contained information is especially sensitive. Developing data dissemination strategies for these data types is an area of current research.

c) Database privacy technologies

Database privacy technologies are required when analysts are restricted from viewing all information contained on the database. The meaning of database privacy is largely dependent on the context where this concept is being used. In co-operative market analysis, it is understood as keeping private the databases owned by the various collaborating corporations. In healthcare, both of the above requirements may be implicit: patients must keep their privacy and the medical records should not be transferred from a hospital to, say, an insurance company. In official statistics, it normally refers to the privacy of the respondents to which the database records correspond. In the context of interactively queryable databases and, in particular, Internet search engines, the most rapidly growing concern is the privacy of the queries submitted by users (especially after scandals like the August 2006 disclosure by the AOL search engine of 36 million queries made by 657000 users). Thus, what makes the difference is whose privacy is being sought.

The last remark motivates splitting database privacy in the following three dimensions:

1. **Owner privacy** is about two or more autonomous entities being able to compute queries across their databases in such a way that only the results of the query are revealed.
2. **Respondent privacy** is about preventing re-identification of the respondents (e.g., individuals like patients or organizations like enterprises) to which the records of a database correspond. Usually respondent privacy becomes an issue only when the database is to be made available by the data collector (hospital or national statistical office) to third parties, like researchers or the public at large. This type of privacy is relevant to the discussion on remote data access, mentioned in b) above.
3. **User privacy** is about guaranteeing the privacy of queries to interactive databases, in order to prevent user profiling and re-identification.

The technologies to deal with the above three privacy dimensions have evolved in a fairly independent way within research communities with surprisingly little interaction.

(i) Privacy-preserving data mining (data owner’s privacy)

Owner privacy is the primary though not the only goal of privacy-preserving data mining (PPDM). PPDM has become increasingly popular because it allows sharing sensitive data for analysis purposes. It consists of techniques for modifying the original data in such a way that the private data and/or knowledge (i.e. the rules) remain private even after the mining process. PPDM may provide also respondent privacy as a by-product.

PPDM based on random perturbation was introduced by Agrawal and Srikant (2000) in the database community. This type of PPDM is largely based on statistical disclosure control (discussed below). Independently and simultaneously, PPDM based on secure multi-party computation was introduced by Lindell and Pinkas (2000) in the cryptographic community. We next discuss some specifics of PPDM for data hiding and for knowledge hiding.

**PPDM for data hiding**

The approach can be oriented to countering attribute disclosure (protecting values of confidential attributes) or identity disclosure (protecting against re-identification).

Random perturbation PPDM for data hiding uses the statistical disclosure control methods (see below). Cryptographic PPDM for data hiding uses secure multi-party computation for a wide range of clustering and other data mining algorithms in distributed environments, where the data are partitioned across multiple parties. Partitioning can be vertical (each party holds the projection of all records on a different subset of attributes), horizontal (each party holds a subset of the records, but each record contains all attributes) or mixed (each party holds a subset of the records projected on a different subset of attributes). For example, a secure scalar product protocol based on cryptographic primitives is applied in privacy preserving k-means clustering over a vertically distributed dataset in Vaidya and Clifton (2000).

Using secure multi-party computation protocols which are based on cryptography or sharing intermediate results often requires changing or adapting the data mining algorithms. Hence, each cryptographic PPDM protocol is designed for a specific data mining computation and in general it is not valid for other computations. In contrast, random perturbation PPDM is more flexible: a broad range of data mining computations can be performed on the same perturbed data, at the cost of some accuracy loss.

**PPDM for knowledge hiding**

Knowledge hiding or rule hiding (Verykios and Gkoulalas-Divanis, 2008) refers to the process of modifying the original database in such a way that certain sensitive (confidential) rules are no longer inferable, while preserving the data and the non-sensitive rules as much as possible. A classification rule is an expression $r: X \rightarrow C$, where $C$ is a class item (e.g. Credit=Yes in a data set classifying customers between those who are granted credit or not) and $X$ is an itemset (set of values of attributes) containing no class item (e.g. Gender=Female, City=Barcelona). The support of a rule $r: X \rightarrow C$ is the number of records that contain $X$, whereas the confidence of the rule is the proportion of records that contain $C$ among those containing $X$. Rule hiding techniques change the data to decrease the confidence or support of sensitive rules to less than the minimum confidence or support threshold required for a rule to be inferred. Data perturbation, data reconstruction, data reduction and data blocking are some principles that have been proposed to implement rule hiding.

(ii) Private information retrieval (user’s privacy)
In proxy-based approaches, the user sends her query to a proxy, who centralizes and submits queries from several users to the database. Thus, it is assumed that the user knows the physical address of the sought item, which might be too strong an assumption in many practical situations. Keyword PIR (Chor et al. 1995) is a more flexible form of PIR: the user can submit a query consisting of a keyword and no modification in the structure of the database is needed.

The PIR protocols in the above cryptographic sense have two fundamental shortcomings which hinder their practical deployment:

1. The database is assumed to contain $n$ items and PIR protocols attempt to guarantee maximum privacy, that is, maximum server uncertainty on the index $i$ of the record retrieved by the user. Thus, the computational complexity of such PIR protocols is $O(n)$. Intuitively, all records in the database must be "touched"; otherwise, the server would be able to rule out some of the records when trying to discover $i$. For large databases, $O(n)$ computational cost is unaffordable.

2. It is assumed that the database server co-operates in the PIR protocol. However, it is the user who is interested in her own privacy, whereas the motivation for the database server is dubious. Actually, PIR is likely to be unattractive to most companies running queryable databases, as it limits their profiling ability. This probably explains why no real instances of PIR-enabled databases can be mentioned.

If one wishes to run PIR against a search engine, there is another shortcoming beyond the lack of server co-operation: the database cannot be modeled as a vector in which the user can be assumed to know the physical location of the keyword sought. Even keyword PIR does not really fit, as it still assumes a mapping between individual keywords and physical addresses (in fact, each keyword is used as an alias of a physical address). A search engine allowing only searches of individual keywords stored in this way would be much more limited than real engines like Google and Yahoo. In view of the above, several relaxations of PIR have been proposed. These fall into two categories:

- **Standalone relaxations.** In Domingo-Ferrer (2009), a system named Goopir is proposed in which a user masks her target query by OR-ing its keywords with $k$ fake keyword sets of similar frequency and then submits the resulting masked query to a search engine or database (which does not need to be aware of Goopir, let alone co-operate with it). Then Goopir locally extracts the sub-set of query results relevant to the target query. Goopir does not provide strict PIR, because the database knows that the target query is one of the $k$ OR-ed keyword sets in the masked query. TrackMeNot (Howe and Nissenbaum, 2009) is another practical system based on a different principle: rather than submitting a single masked query for each actual query like Goopir, a browser extension installed on the user's computer hides the user's actual queries in a cloud of automatic "ghost" queries submitted to popular search engines at different time intervals. Again, this is not strict PIR, because the target query is one of the submitted ones.

- **Multi-party relaxations.** The previous standalone approaches rely on fake queries and this can be viewed as a weakness: it is not easy to generate fake queries that look real and, besides, overloading the search engines/databases with fake queries clearly degrades performance. Multi-party relaxations avoid fake queries by allowing one user to use the help of other entities either for anonymous routing or for cross-submission:

  - An onion-routing system like Tor (Tor Project, 2014) is not intended to offer query profile privacy because it only provides anonymity at the transport level; however, if complemented with the Torbutton component, Tor can offer application-level security (e.g. by blocking cookies) and it can be used for anonymous query submission.

  - In proxy-based approaches, the user sends her query to a proxy, who centralizes and submits queries from several users to the database/search engine. While using the proxy prevents the database/search engine from profiling the user, the proxy itself can profile the user and this is a weakness.

  - Another principle is peer-to-peer: a user can submit queries originated by peers and conversely. In this way, peers use each other's queries as noise. The advantage here is that this noise consists of real queries by other peers, rather than fake queries as in standalone systems. The game-theoretic analysis in Domingo-Ferrer (2012) shows that cross-submission among peers is a rational behavior. Proposals exploiting this approach include Crowds (Reiter and Rubin, 1998) and more recently by Domingo-Ferrer (2009).

*(iii) Remote access statistical disclosure control*

In the case of remote access to a database as discussed in b (iii), a simple model for a remote server is:

1. An analyst submits a query (in this case, requests a frequency table), via the internet, to the remote server.
2. The remote server processes the analyst's query on the linked micro-data.
3. The outputs (e.g. counts) are modified or restricted in order to ensure the risk of disclosure is acceptably low.
4. The remote server sends the modified output, via the internet, to the analyst.

The aim of statistical disclosure control (i.e. Step 3) is to ensure that the risk of disclosure from releasing output via the remote server is acceptably small. Seemingly harmless multivariate analyses such as linear regressions can be manipulated to reveal information on the individual level (see for example, Bleninger et al. 2011 or Drechsler et al. 2014). Identity disclosure would occur if an attacker was able to infer that a particular person belongs to the database; membership of a database could be sensitive (e.g. if the database contained records about patients suffering a particular medical condition). If an attacker knows a person is on the database, attribute disclosure occurs if an attacker is able to infer the characteristics of the person from the database.

As mentioned in step 3, there are two main SDC principles for queryable database protection:

- **Query perturbation:** Perturbation (noise addition) can be applied to the microdata records on which queries are computed (input perturbation) or to the query result after computing it on the original data (output perturbation).
- **Query restriction:** The database refuses to answer certain queries.

Query restriction is the right approach if the user does require deterministically correct answers and these answers have to be exact (i.e. a number). Exact answers may be very disclosive, so it may be necessary to refuse answering certain queries at some stage. A common criterion to decide whether a query can be answered is query set size control: the answer to a query is refused if this query together with the previously answered ones isolates too small a set of records. The main problem of query restriction are: i) the computational burden to keep track of previous queries; ii) collusion attacks can circumvent the query limit.

d) Privacy of estimates and microdata
Measuring the risks to privacy when a statistical agency releases estimates or micro-data has been given considerable attention in the literature (for a review see Hunderpool et al. 2012).

(i) Microdata

Microdata releases allow analysts to view the data values for individual records. Even though names and address information is removed from the microdata before its release, the risk of disclosure remains. Disclosure risks for microdata releases include:

- The risk of identifying a person by linking a record on the microdata to an administrative file that contains name and address, via a set of quasi-identifiers (e.g. age, sex, suburb, occupation)
- Spontaneous recognition, i.e., the direct identification of individuals based on a set of quasi-identifiers provided in the dataset without using any external information, such as administrative file.

Let \( N_q \) be the number of people in the population with quasi-identifier \( q \). The above disclosure risks are often related to \( 1/N_q \).

A non-perturbative method to manage the disclosure risks is to reduce the detail in of data items that could be used as quasi-identifiers. The less detail provided on the microdata, the higher the value of \( N_q \), and the lower the disclosure risk. For continuous variables common strategies include discretizing (e.g. reporting only 5 year age groups instead of the exact ages) Another strategy is top coding or replacing the values above a certain threshold with the value of the threshold (e.g. income values above 100,000 EUR might be reported as “100,000+”). For categorical variables a common strategy is to collapse sparsely populated categories.

Common perturbation methods include: swapping, adding random noise (e.g. PRAM for categorical data) and micro-aggregation. Perturbative methods can be applied to potential quasi-identifiers, in order to reduce the risk of identification. Perturbative methods can also be applied to sensitive variables, and thereby reduce the risk of attribute disclosure.

An alternative to traditional perturbation is to replace original data with synthetic data. Under the partially synthetic data approach only the sensitive records and/or records that could be used for re-identification are replaced with synthetic values. The synthesis can range from synthesizing only some records for a single variable, for example all incomes for individuals with an income above a given threshold, to synthesizing all variables. When making inference with synthetic data, special formulas are required (Reiter, 2003; Reiter, 2005a). Since true values remain in the released data careful disclosure risk evaluations are still necessary. Methods to evaluate these risks haven been suggested in Drechsler and Reiter (2008) and Reiter and Mitra (2009). Examples of this approach include: in the U.S., the Survey of Consumer Finances by replaced monetary values at high risk of disclosure with multiple imputations (Kennickell, 1997); On the Map (Machanavajjhala et al., 2008), illustrating commuting patterns (i.e., where people live and work) for the entire United States via maps available to the public on the Web (http://lehdmap.did.census.gov/); the Survey of Income and Program Participation (SIPP) (Abowd et al., 2006), the Census Bureau’s synthetic version of the Longitudinal Business Database (Kinney et al., 2011) and in Germany the Institute for Employment Research released a synthetic version its establishment survey in 2011 (Drechsler, 2012).

(ii) Tabular estimates

Below the risks associated of releasing tabular estimates are considered. Tables may contain counts (e.g. counts of people) or sums of continuous data (e.g. person income).

Tabular frequency counts

Counts are often presented in a table which have multiple dimensions (e.g. number of employed people by postcode, age and sex). The dimensions of the table (e.g. postcode, age and sex) could identify a single person as the only contributor to a cell and thereby reveal attributes of the person (e.g. employment status), in the same way as for a micro-data release. Count estimates are at risk from differencing attacks. A differencing attack involves taking the difference between two cell counts, where the two counts are the same except that a small number of people are excluded from one of the counts. The difference reveals the count for those people who were excluded.

A non-perturbative protection measure for count estimates is to restrict tables which have dimensions that could identify a single person. Another approach is to perturb the cell counts- even if the dimensions of a table identify a single person, perturbation can be designed to protect the person’s attributes from being disclosed.

Tabular magnitude estimates

An example of a magnitude estimate is total business income by postcode, age and sex. Identify disclosure, via a table’s dimensions, is possible in the same way as (ii) above. Attribute disclosure is often defined in terms of how close (e.g. percentage difference) an attacker is able to infer a person’s attribute.

e) Evaluation of different privacy approaches

When an organization applies statistical disclosure control to microdata, or estimates derived from them, there is an implicit trade-off between the disclosure risk and the utility. Utility is a measure of how well statistical output from the micro-data meets the needs of the analyst. Utility is often measured by the difference between the statistical output before and after disclosure control methods have been applied, where only the latter is released to the analyst. It is clearly undesirable if statistical disclosure control seriously affects an analysts’ conclusions. Utility can also be measured in terms of whether an analyst can obtain the desired output at all. For example, there is no constraint on what output an analyst can produce if the microdata are available on CD Rom; however, there is a restrictive set of analytical outputs available from a remote server.

Disclosure risk depends upon the context, or “scenario” in the words of Hunderpool et. al (2012), in which the attack is assumed to occur. A scenario describes the:

1. information available to identify the target unit (i.e. matching key variables). This could be in the form of prior knowledge or in the form of administrative data containing name and addresses;
2. the method by which an attacker attempts to disclose sensitive information about the target.

Hunderpool et. al (2012) give several examples of scenarios. While a wide range of attacks are possible, a pragmatic approach for an agency to take is to focus on the scenarios which are likely to have the greatest disclosure risk. Below a summary of the advantages and disadvantages of the different approaches to privacy are discussed. This will be helpful in the next chapter, where the suitability of the different approaches to Big Data will be discussed.

(i) Utility and disclosure risk for queryable databases

From a utility/risk perspective, some advantages of remote analysis are:

- Although the statistical output is modified, it is based on the real microdata. This means complex relationships in the microdata can be essentially retained.
- The degree to which a particular output is modified can depend upon the output itself. For example, estimates at a broad level may require less modification than estimates at a fine, or small area, level.
- Since an analyst is restricted from viewing the attributes of any record, less protective measures may be needed than would otherwise be the case. This means that a small number, perhaps only one, record can be attacked at a time.
- The impact of the modifications on the output can be broadly indicated to the analyst or incorporated into the output seamlessly.
- It can process multiple queries in real time.
- Submitted queries can be logged and audited. If a disclosure attempt is detected and confirmed, the agency can revoke the analyst's access and/or take legal action.

There are some disadvantages of remote analysis.

- Some statistical outputs may be aggregated (e.g. record-level residual plots may be replaced with parallel box plots), perturbed (e.g. random noise added to regression coefficients), or restricted altogether.
- The analyst may be restricted to submit programming code from specific statistical packages (e.g. SAS), or may be restricted to using a menu-driven interface.
- Also, remote analysis may take longer than if the micro-data were available on the analyst's personal computer.

(ii) Utility and disclosure risk in microdata disclosure control

The advantages and disadvantages of microdata disclosure control are more or less the inverse of the advantages and disadvantages of the queryable databases. Advantages include:

- All information is available at the microlevel and there are no restrictions on the types of analyses that can be applied.
- The analyst can use any software for his or her research.
- The analyst doesn’t have to wait before the research output is checked for any disclosure violations.
- Organizations don’t need to worry that a clever combination of different research queries might reveal more information than anticipated.

Disadvantages include:

- If perturbative methods have been used to protect the data, there is no guarantee that the results based on the protected data are similar to the results that would have been obtained based on the original data.
- All protection mechanisms are applied to ensure that no disclosure occurs even if the full microdata is available. This usually requires a higher amount of perturbation than would be necessary if only a specific output needs to be protected.
- It is often difficult for the user to get an idea how much his or her analysis has been affected by the protection mechanisms that have been applied prior to the release.
- If a risk of disclosure is detected after the data have been disseminated it is nearly impossible to ensure that all released files are deleted.

(iii) Utility and disclosure risk for tabular data

Typically, the disclosure risks associated with tabular data are typically much lower than for microdata because the available quasi-identifiers in the case of the former have less detail. Statistical organizations typically have basic rules for the publication/release of a particular tabular estimates (see Anna Oganian and Josep Domingo-Ferrera 2003). These may include the “Rule of Two” which means that no non-zero cell count can be based on fewer than two people. Often users are not satisfied with the limited set of tables that a statistical organization is able to publish- this naturally means the statistical organization will also have to allow another type of data access.

f) Privacy software

(i) Remote server

Table Builder is a remote server developed by the Australian Bureau of Statistics that allows users access to highly detailed micro-data in order to:

- Construct user-defined tables (e.g. state, age, sex)
- Calculate means, medians and quantiles for continuous variables (e.g. income) and their associated relative standard errors
- Create custom ranges from continuous variables
- Create graphs using your customised tables
- Save or export your tables as CSV, Excel or SDMX files

DataAnalyser is a remote server developed by the Australian Bureau of Statistics (Chipperfield, 2014 and Chipperfield and O'Keefe, 2014) that allows users access to highly detailed microdata in order to undertake analyses using a menu driven user interface. DataAnalyser allows users to transform and manipulate variables and to produce basic exploratory data analysis, summary tables and regressions analysis (e.g. linear (robust), logistic, probit and multinomial). Condentialised outputs can be viewed on-screen or downloaded to the user's own computer. For more information see http://www.abs.gov.au/websitedbs/D3310114.nsf/home/About+DataAnalyser. The confidentiality method underlying Data Analyser is described by Chipperfield and O'Keefe (2014)
Microdata dissemination

The Special Uniques Detection Algorithm (SUDA) (Elliot, Manning and Ford, 2002) is an algorithm to locate 'risky' records with categorical variables. First, SUDA finds the different sets of variables on which a record is unique. Second, SUDA ranks the 'risk' of each record in terms of the number and distribution of these sets. For example, SUDA will find "Special unique records" which are unique on an "unusual" and small number of attributes (e.g. 16 year-old widow).

(iii) Statistical disclosure control software

- ARGUS is a software program designed to protect statistical tables http://neon.vb.cbs.nl/casc/tau.htm
- µ-ARGUS is a software program designed to create safe micro-data files http://neon.vb.cbs.nl/casc/mu.htm
- sdcMicro based on R : http://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf

2 References


Top editors in past week

None