Challenges for statistical disclosure control in a world with big data and open data

de Wolf, Peter-Paul and Zeelenberg, Kees

Statistics Netherlands, Statistics Netherlands

30 July 2015

Challenges for statistical disclosure control in a world with big data and open data

Peter-Paul de Wolf*
Statistics Netherlands, The Hague, Netherlands – pp.dewolf@cbs.nl

Kees Zeelenberg*
Statistics Netherlands, The Hague, Netherlands – k.zeelenberg@cbs.nl

Abstract

National statistical institutes (NSIs) produce tables and microdata files; these data are typically checked for unwanted disclosure of data of individual persons and enterprises. In recent years many other data sources, in the form of open data and big data, have become available, for the general public, for researchers, and for and from data collectors and providers other than NSIs. This poses new problems for statistical disclosure control. We discuss several of these problems, such as should NSIs protect their tables and microdata files against linking to other datasets so as to prevent detailed profiling of their respondents, should we entertain different disclosure scenarios and switch to other disclosure-control methods?

Keywords: disclosure control; open data; big data.

1. Introduction

Traditionally, statistical disclosure control by national statistical institutes (NSIs) has focused on tables and microdata collected and produced by the NSI itself. In the last decade NSIs have increasingly also been using administrative data, collected by government agencies; and usually, the individual administrative data have not been published and have not been used by other government agencies or private companies.

However, in recent years this context of an almost monopoly for the NSIs has changed. Governments have been actively pursuing an *open-data policy* with easy access to government data, not only as a service to the general public, but also with the express purpose of stimulating the market for data. NSIs are also beginning to offer their tabular data as open data, by providing a public and easy way to download these data in an automatic way. This market for data has been further stimulated by the advent of *big data*, data that come in large amounts, sometimes in millions or even billions of records per day; some examples are telephone records, traffic data, and banking transactions. These big data have made it possible for private data companies to assemble databases with very detailed information on individuals and enterprises, with data coming from many sources, and linked sometimes deterministically but often also probabilistically.

This abundance of data poses new problems for statistical disclosure control, both strategic and ethical problems as well as methodological problems. For example, should NSIs, when protecting tables or data files against disclosure, take into account the possibility, or even the absolute certain fact, that government agencies and private companies will use the NSI data to enrich their own databases and so get to know more about their citizens or customers? And what if these enriched data are used to profile citizens so that they may come under suspicion or are denied access to certain services such as loans? And should we use another methodological paradigm than we have used so far? Should we consider different disclosure scenarios? Does the changing attitude towards privacy influence the way we should treat our published data? How much existing as well as future data should we take into account when assessing disclosure risks?

2. Context of privacy and disclosure control

General rules for privacy are provided in the European Union by the *Data Protection Directive* of 1994 (EU, 1994). On the other hand, the USA does not have a general system of rules for privacy (LLB, 2015). Besides these general rules for privacy, NSIs, also those in the USA, often have explicit commitments to privacy of respondents, not only persons and households but also enterprises. As basis for our discussion we will refer to the relevant principle from the Code of Practice for European Statistics:

“**The privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes are absolutely guaranteed.**

INDICATORS

5.1: Statistical confidentiality is guaranteed in law.

5.2: Staff sign legal confidentiality commitments on appointment.

5.3: Penalties are prescribed for any wilful breaches of statistical confidentiality.

5.4: Guidelines and instructions are provided to staff on the protection of statistical confidentiality in the production and dissemination processes. The confidentiality policy is made known to the public.

5.5: Physical, technological and organisational provisions are in place to protect the security and integrity of statistical databases.

5.6: Strict protocols apply to external users accessing statistical microdata for research purposes.”

3. Open data

3.1 What are open data?

Open data are data that can be “*freely used, reused and redistributed by anyone*” (http://opengovernmentdata.org). Very often these open data come from government or government-sponsored institutions, and they have been made open because the releasing institution wants to increase transparency of its operations and outcomes and of citizen participation. In the Netherlands, local governments (municipalities) in particular want to provide their citizens with detailed information about their activities, which in 2015 have increased considerably due to decentralization of care for elders, youngsters and people with disabilities. At present there are over 3000 open-data files, both on the national and the local level, and include for example data on traffic intensities, environmental measurements, laws, schools, and locations of houses and structures (http://data.overheid.nl/). In other countries there are similar initiatives, for example in the USA: http://www.data.gov/open-gov/. Also, Statistics Netherlands has made all its published data available in an open-data format (http://www.cbs.nl/nl-NL/menu/cijfers/statline/open-data/default.htm).

Apart from transparency and citizenship, governments also want to stimulate the private data sector; see for example the Open Data Institute (ODI; http://opendatainstitute.org) in the UK and the general framework in the EU for the re-use of public data (EU, 2003).

3.2 Open data, privacy and disclosure control

Both NSIs and other data providers supply open data in tabular form as well as in public-use micro data files. All data providers are bound to the privacy rules of national and international laws, and so they are obliged, and usually will have, safeguarded the privacy of the persons in their tables and micro data, both the publications as such and in combination. However the combination of several data sources from different providers may pose additional opportunities for disclosure, intentionally or unintentionally. There are various ways for an NSI to deal with these disclosure risks:

1. The NSI may protect its open data taking into account other data sources. This puts the burden of disclosure control at the NSI and would severely limit the details and the usefulness of its data. But this would become less of a problem if the NSI would come to an agreement with

other data providers, in particular public providers, about the level of detail to be published and about disclosure-control techniques to be used.

2. The NSI may act as an open-data provider for all public data, i.e. data from government or government-controlled institutions, and protects all public data in connection with each other. NSIs have a long tradition and extensive experience as publishers of public data and with disclosure-control methods. It would be an extension of the present position and status of NSIs, but it would certainly be appropriate and in line with a digital information strategy for official statistics. In this case also, the NSI would have to come to an agreement with its partners for whom it acts as data provider, about the level of detail to be published and about disclosure-control techniques to be used.
3. The NSI may add legal conditions to the use of its open data that prohibit disclosure of individual data also when it is achieved by combining the NSI data with other public or private data sources.

In practice a combination of these three measures may prove to be the way to go. For public data, a combination of the first two would be useful, and for limiting disclosure risks of combining public data with private data, the third measure would be useful.

Especially for tabular data, disclosure control is usually done by suppressing (Hundepool et al, 2014). Many users, in particular local governments directly in contact with their citizens, however, prefer a common information structure for all units, for example neighborhoods, distinguished in tables. Coarsening of classifications for sensitive variables or rounding of the values may then be alternative, more acceptable, techniques.

4. Big data

There is currently no clear, uniformly accepted definition of Big Data. However, when dealing with Big Data, often the ‘three V’ are mentioned: Volume (amount of data), Velocity (speed of data in and out) and Variety (range of data types and sources). Each V poses several questions and issues related to Statistical Disclosure Control:

Volume

How will NSIs deal with huge amounts of data? In general they will still publish aggregated information. In that case the current SDC techniques might still be applicable. However, when Big Data (or excerpts from them) are released as microdata, the current SDC techniques no longer apply: identity disclosure is almost certain. Then methods producing synthetic data may be preferable: use the Big Data sets to estimate a model and use that model to generate a synthetic dataset that resembles the original Big Data. Or other techniques that mask the true values of sensitive variables: just create enough uncertainty about the exact values to make it “less sensitive”.

Velocity

When data become available more quickly, the processing of the data also needs to be done in less time. The current SDC techniques can be time consuming when the underlying datasets increase in size and number. Streaming data might lead to streaming statistics, but then we should be able to protect those statistics “real-time”.

Variety

With Big Data you might end up with unstructured data, distributed over different ‘places’, indirect observations (events instead of units), unclear underlying population, selectivity, etc. This does not only influence the SDC methods that can be used, but just as well the disclosure risk scenarios that need to be taken into account.

Current risk models need a clear definition of the underlying population: uncertainty about individual information is usually introduced by making sure that there are enough look-alike units in the underlying population. But when the underlying population is not known, this is not a valid option. Uncertainty should then be attained by introducing uncertainty on the sensitive information directly.

5. Discussion

Traditionally, disclosure control is more focused on identity disclosure: the risk of identifying a unit in statistical output and consequently deriving sensitive information about that unit. SDC methods aim at reducing the risk of identifying units. As a second measure to prevent disclosure, other methods were developed that mask the sensitive variables. In some countries the laws specifically state that it should not be possible to identify units in publications, not even by those units themselves. With huge amounts of data, be it from a large collection of Open Data or from Big Data source(s) directly, identity disclosure might become too easy. there is just too much (indirect) identifying information available or derivable. Maybe disclosure control should focus more on attribute disclosure: the risk of deriving (sensitive) information about an individual unit. Masking sensitive data for example might lead to identification of an individual unit, but with enough uncertainty about the exact value of the sensitive attribute.

Now Open Data is providing a huge amount of (related) information, an interesting question arises: is privacy protection a legal issue only or is it an ethical issue as well? Suppose that an NSI published data and that data can be used to derive some sensitive information stemming from a source that was not published by that NSI. Should an NSI try to prevent this from happening? The NSI did not publish that sensitive information as its own publication. Hence legally the NSI does not need to take this into account. But how about ethics, moral?

Computer power has increased tremendously over the last 10 to 20 years. Together with more and more data sets becoming available (Open Data), this means that an “attacker” has more power and more possibilities to try to breach confidentiality in publications of NSIs. Disclosure control is always related to certain “attacker scenarios”. The traditional attacker scenarios are becoming a bit obsolete and new attacker scenarios might need to be considered.

Not all issues can be solved by inventing new methodology and coding new software. A general, public discussion is needed on the notion of privacy. This may lead to adjustment of laws (or the interpretations thereof) concerning privacy. Current laws are usually based on the ‘old’ practice of precisely defined populations, sample surveys and ‘old’ technological possibilities.

A general, public discussion is needed to place privacy and confidentiality in the right perspective. The general public is keen on complaining that institutes should not breach their privacy, whilst at the same time they publish very detailed information about themselves via social media. This shows that the notion of privacy changes over time: things that were considered sensitive several years ago, are no longer considered to be sensitive now; or the other way around. Knowing about expected routes for convoys of chemical or nuclear trucks might be considered to be more sensitive since September 2001.

References

EU (European Union) (1994), Data Protection Directive. http://ec.europa.eu/justice/data-protection/

EU (2003), Directive on the re-use of public sector information. http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information

Hundepool, Anco & Josep Domingo-Ferrer & Luisa Franconi & Sarah Giessing & Eric Schulte Nordholt & Keith Spicer & Peter-Paul de Wolf (2012), Statistical Disclosure Control. Wiley, New York.

Jolly, I. (2015), Data protection in United States: overview. http://uk.practicallaw.com/6-502-0467#