24 Jun 2016

Statistical Disclosure Control[1]

What is Statistical Disclosure Control (SDC)?

Statistical Disclosure Control (SDC) refers to the application of statistical techniques to microdata (i.e. data at the individual entity level) so that these processed datasets can be made available to researchers without undue leakage of information on any one individual.

Why the need for SDC?

The granularity afforded by microdata presents the opportunity to gain deeper insights for policy research and review. However, releasing microdata as-is severely compromises the confidentiality of individuals. SDC seeks to transform microdata in a way that protects the confidentiality of individuals while minimising information loss to the researcher.

What are some measures of disclosure risk?

There have been several different approaches to measuring disclosure risk. The most common and simplest measures are based on the idea that records with unique combinations of key variable values have higher risks of re-identification.

- ***Does the dataset satisfy k-anonymity?*** A dataset satisfies *k*-anonymity if each distinct pattern of key variables is possessed by at least *k* records in the sample. This means that knowing the key variable values for an individual does not allow an adversary to identify which record belongs to the individual: there will be *k* of them to choose from.

  The higher the value of *k*, the lower the disclosure risk. A typical practice is to set *k* = 3. As an example, the dataset represented by Table 1 below satisfies 3-anonymity but not 4-anonymity.

- ***Does the dataset satisfy l-diversity?*** A dataset satisfies *l*-diversity if observations with the same pattern of key variables contains at least *l* "well-represented" values for the sensitive variables. This is a stronger notion of privacy than *k*-anonymity.

  As an example, the first 3 records in Table 1 are 2-diverse while the last 3-records are 1-diverse. It is clear that 1-diversity leaks information: an adversary would know that anyone who was female and in her 30s would have cancer.

Table 1

	Key variables		Sensitive variable
	Gender	Age group	Medical condition
1	Male	30s	Cancer

[1] This section is largely based on Templ et al., "Introduction to statistical disclosure control (SDC)." International Household Survey Network (IHSN) Working Paper No 007, Aug 2014. http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf.

2	Male	30s	Heart Disease
3	Male	30s	Heart Disease
4	Female	20s	Cancer
5	Female	20s	Cancer
6	Female	20s	Cancer

Another angle to approach disclosure risk is by estimating the expected number of reidentifications for the dataset.

Common SDC techniques

There are 3 broad kinds of SDC techniques:

a. Non-perturbative techniques, which suppress or reduce the detail without altering the original data;
b. Perturbative techniques, which distort the original data before release; and
c. Synthetic data generation, which creates fictitious datasets that preserve certain statistics or relationships in the original dataset.

Below, we describe basic non-perturbative and perturbative techniques that are easy to administer.

Non-perturbative techniques

- **Recoding.** Recoding refers to the collapsing several categories into one with a higher frequency count and less information. The table below shows how recoding can be used for both categorical and continuous variables.

Level of schooling	Income
Primary	$4,050
Secondary	$4,070
Secondary	$10,250
Tertiary	$6,700
Postgraduate	$12,500
Postgraduate	$6,300

Level of schooling	Income
Secondary & below	$4,000-$4,999
Secondary & below	$4,000-$4,999
Secondary & below	> $10,000
Tertiary & above	$6,000-$6,999
Tertiary & above	> $10,000
Tertiary & above	$6,000-$6,999

- **Local suppression.** Local suppression adds missing values to certain key variables so as to help achieve *k*-anonymity.

Perturbative techniques

- **Micro-aggregation.** Typically applied to continuous variables, micro-aggregation partitions records into groups, then assigns an aggregate value (typically the mean) to each variable in the group. The table below shows an example of micro-aggregation to achieve 2-anonymity.

Income
$4,050
$4,070
$6,300
$6,700
$10,500
$12,500

Income
$4,060
$4,060
$6,500
$6,500
$11,500
$11,500

- **Adding noise.** For continuous variables, a randomised number can be added or multiplied to the original values to protect data from exact matching with external files.

Synthetic data generation

There are packages in R which provide routines to generate synthetic versions of datasets. Nevertheless, synthetic datasets can only preserve the relationships which the analyst specifies beforehand, and thus results obtained from them may not be as reliable as those obtained from datasets more closely resembling the original dataset.

Differential Privacy

What is differential privacy?

Differential privacy is a *property* of a *protocol (or mechanism)* which runs on some dataset and produces an output.[2] It is neither an algorithm in itself, nor is it the property of a database on its own.

Intuitively, a protocol is **differentially private** if it produces very similar outputs on databases which differ in or by just one record. Since the results are similar, it is difficult for an adversary to determine the contribution of that one change to the result, and hence little information on that record is lost.

Please see the *Technical Annex* for a formal definition of differential privacy, along with examples of protocols which are and are not differentially private.

Why the need for differential privacy?

Differential privacy was first conceptualised in 2006 by Cynthia Dwork, a computer scientist with Microsoft Research. Differential privacy addresses the issue of releasing statistical properties of a database without revealing information about any one particular value. In particular, "differential privacy promises to protect individuals from any *additional* harm that they might face due to their data being in a private database x that they would have faced had their data not been part of x."[3]

Differential privacy is not the first framework which tries to address this problem. However, many of the earlier frameworks (e.g. *k*-anonymity) are known to be susceptible to attacks.[4]

How can differential privacy be achieved?

Finding differentially private mechanisms is an active field of research. Fundamentally, any differentially private mechanism must have a random component to it.[5] The most common method of achieving differential privacy is by adding a controlled amount of random noise to the query result (e.g. Laplace mechanism). Other more complicated algorithms (e.g. exponential mechanism) exist as well.

Limits of differential privacy

[2] Blum, "An brief tour of Differential Privacy." http://www.cs.cmu.edu/afs/cs/academic/class/15859m-s11/www/lectures/lect0420.pdf.

[3] p20, Dwork and Roth, *The Algorithmic Foundations of Differential Privacy*. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.

[4] Tu, "Lecture 20: Introduction to Differential Privacy." http://people.eecs.berkeley.edu/~sltu/writeups/6885-lec20-b.pdf. See Sweeney, "*k*-Anonymity: A model for protecting privacy" (https://epic.org/privacy/reidentification/Sweeney_Article.pdf) for possible attacks on *k*-anonymity.

[5] p16, Dwork and Roth.

Differential privacy is still subject to the fundamental tradeoff between accuracy and privacy, which can be a problem when training complex machine learning models.[6] For example, in exploring the tradeoff between privacy and utility in using differentially-private algorithms to guide dosage levels in warfarin therapy[7], Fredrikson et al.[8] found that using these algorithms at ε values protecting patient privacy resulted in unacceptable levels of risk of negative patient outcomes.

Differential privacy in practice

Differential privacy is not widely used despite much active research in academic research circles.

- In 2008, a research team from Cornell University used a differentially private process to generate a synthetic dataset[9] of commuting patterns from US Census Bureau data.[10]
- Google has been running its RAPPOR (Randomised Aggregatable Privacy-Preserving Ordinal Response) project since 2014. RAPPOR uses the concept of randomised response "to enable learning statistics about the behaviour of users' software while guaranteeing client privacy."[11]
- In 2016, Apple announced its intention to use differential privacy in iOS 10.

Differentially private mechanisms and statistical disclosure control

Differentially private mechanisms differ fundamentally from traditional statistical disclosure control techniques in that they generally apply to different set-ups. The risk criterions against which mechanisms are evaluated are different as well, and the 2 paradigms are not easily reconciled.[12]

- **Differential privacy:** Raw data resides in a database. The analyst does not have access to the raw data itself. A differentially private algorithm prescribes what queries can be executed on the database, as well as the results that these queries return. The goal is to ensure that information on any individual is not leaked through the query results.
- **Statistical disclosure control (SDC):** There is a raw dataset consisting of microdata which an analyst wants access to. Statistical disclosure control processes the raw dataset so that the analyst can have a processed dataset at the same granularity, with minimal information loss but maintaining an acceptable level of reidentification risk.

[6] Green, "What is Differential Privacy?" http://blog.cryptographyengineering.com/2016/06/what-is-differential-privacy.html.

[7] A procedure used to help prevent strokes in patients with a form of irregular heartbeat.

[8] Fredrikson et al., "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing." 23rd USENIX Security Symposium, 20-22 Aug 2014. https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf.

[9] A synthetic dataset is a fictitious dataset which retains some of the statistical properties of the original dataset it was generated from.

[10] Machanavajjhala et al., "Privacy: Theory meets practice on the map." 24th International Conference on Data Engineering, 2008. http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf.

[11] "Learning statistics with privacy, aided by the flip of a coin." Google Security Blog. https://security.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html.

[12] McClure and Reiter, "Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data." Transactions on Data Privacy, 2012. http://www.tdp.cat/issues11/tdp.a093a11.pdf.

There are differentially private mechanisms which can transform a raw dataset into a synthetic dataset. This however does not mean that SDC is irrelevant: there are several other SDC techniques which do not result in the creation of synthetic datasets, but datasets which are closer to the original raw dataset.

Technical Annex

Formal definition of differential privacy

A protocol A acts on a dataset X to produce an output $A(X)$. Fix a small positive parameter ε (ε is sometimes referred to as the *privacy budget*). A is **ε-differentially private** if for any 2 datasets X, X' which differ in or by one record, we have

$$1 - \varepsilon \approx e^{-\varepsilon} \leq \frac{Pr(A(X) = v)}{Pr(A(X') = v)} \leq e^{\varepsilon} \approx 1 + \varepsilon$$

for all all possible outcomes v.[13] The smaller the value of ε, the more privacy afforded to the individual.

Example of a protocol that is differentially private[14]

Consider the following protocol to determine the proportion, p, of students who have cheated in the past year:

> **Step 1:** Flip a coin.
> **Step 2:** If coin is tails, respond truthfully. If not, flip a second coin and respond "Yes" if heads and "No" if tails.
> (Coin flips are not recorded.)

In this protocol, changing the answer for any one student does not yield information, since "Yes" occurs with probability at least 1/4 whether the respondent actually cheated or not. In this way, individual "privacy" is protected. Yet, the value of p can be estimated using simple probability theory (expected proportion of "Yes" answers is $1/4 + p/2$).

Example of a protocol that is *not* differentially private[15]

Suppose we have a database where each record is a pair (*Name*, *Has Diabetes*), where *Has Diabetes* = 1 if the person has diabetes, and is equal to 0 otherwise. Suppose also that the analyst is allowed to issue queries of the form Q_i, which returns the sum of the *Has Diabetes* column for the first i rows.

Name	*Has Diabetes*
A	1
B	0
C	1

[13] Blum.
[14] Adapted from pp15-16, Dwork and Roth.
[15] Adapted from https://en.wikipedia.org/wiki/Differential_privacy.

This protocol is not differentially private: A malicious analyst could issue the queries Q_2 (which returns 1) and Q_3 (which returns 2), subtract the difference and figure out that C has diabetes.[16]

[16] This protocol can be made differentially private if a suitable amount of noise is added to the query result before being returned to the analyst.