Input Privacy: Towards a Logical Framework for Defining and Implementing Official Statistics Use Cases and Scenarios

Monica Scannapieco – Istat

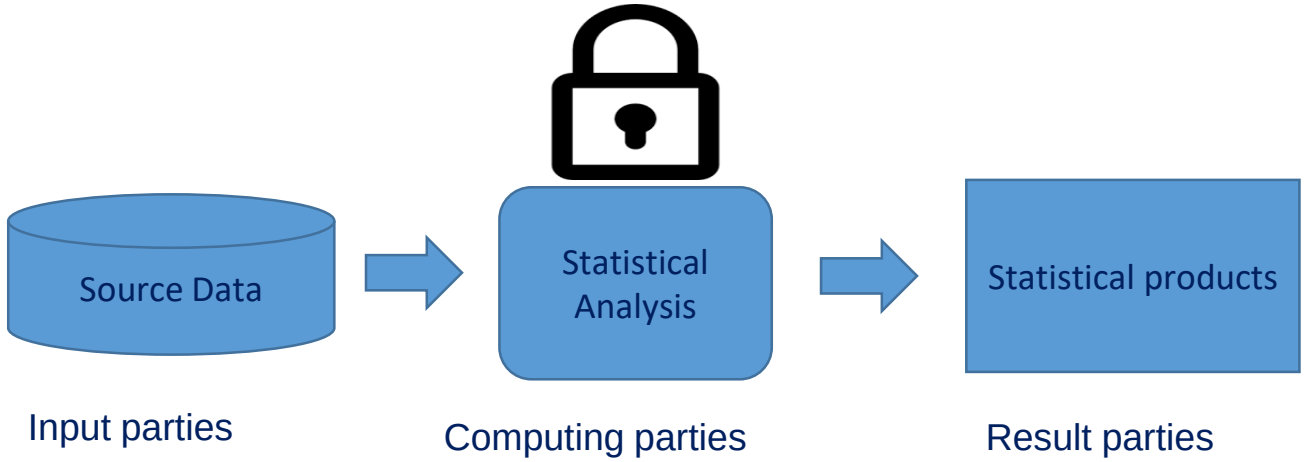
&

IPP Project Team

Outline

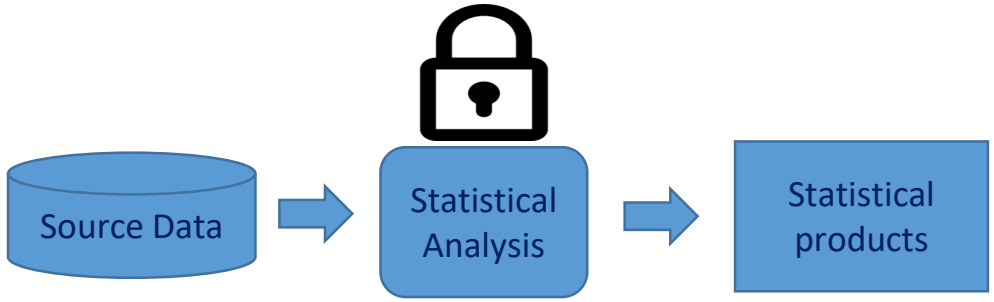
- On defining Input Privacy
- Defining Input Privacy scenarios in Official Statistics: a logical framework
- Instantiating the IPP Template for OS: Private Set Intersection with Analytics
- Conclusions and next steps

Input Privacy: a Definition



- ***Input privacy*** *means that the Computing Party cannot access or derive any input value provided by Input Parties, nor access intermediate values or statistical results available at Result parties during processing of the data (unless the value has been specifically selected for disclosure). [UN Handbook on Privacy-Preserving Computation Techniques, 2019]*

Input Privacy Vs Output Privacy



Input privacy

- Input privacy techniques are based on data «transformations» that preserve source data privacy
- Examples of input privacy techniques: Secure multi-party computation (SMC), homomorphic encryption, trusted execution environment, etc.

Output privacy

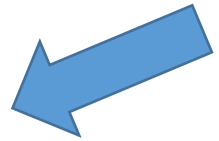
- Output privacy aims at reducing the risk of privacy breaches in the phases of disseminating or exchanging statistical products
- Examples of output privacy techniques or Statistical Disclosure Control techniques: perturbation methods (e.g. differential privacy), non pertubation methods (e.g. local suppression)

Input Privacy: Why

- National Statistical Systems are part of a «data ecosystem»
  - Need to integrate NSIs' data with data of other subjects to build public value
- Investments on producing Big-data based statistics (Trusted Smart Statistics):
  - Several Big data sources are owned by private data holders and there is the need to define protocols enabling NSIs to access to such data in a privacy-preserving way
  - Bucarest Memorandum on Trusted Smart Statistics (October 2018): «*recognise the importance of privacy by design approaches and encourage the ESS to explore the potential of privacy-preserving computation technologies, such as secure multiparty computation, within a wider framework of Trusted Smart Statistics*»

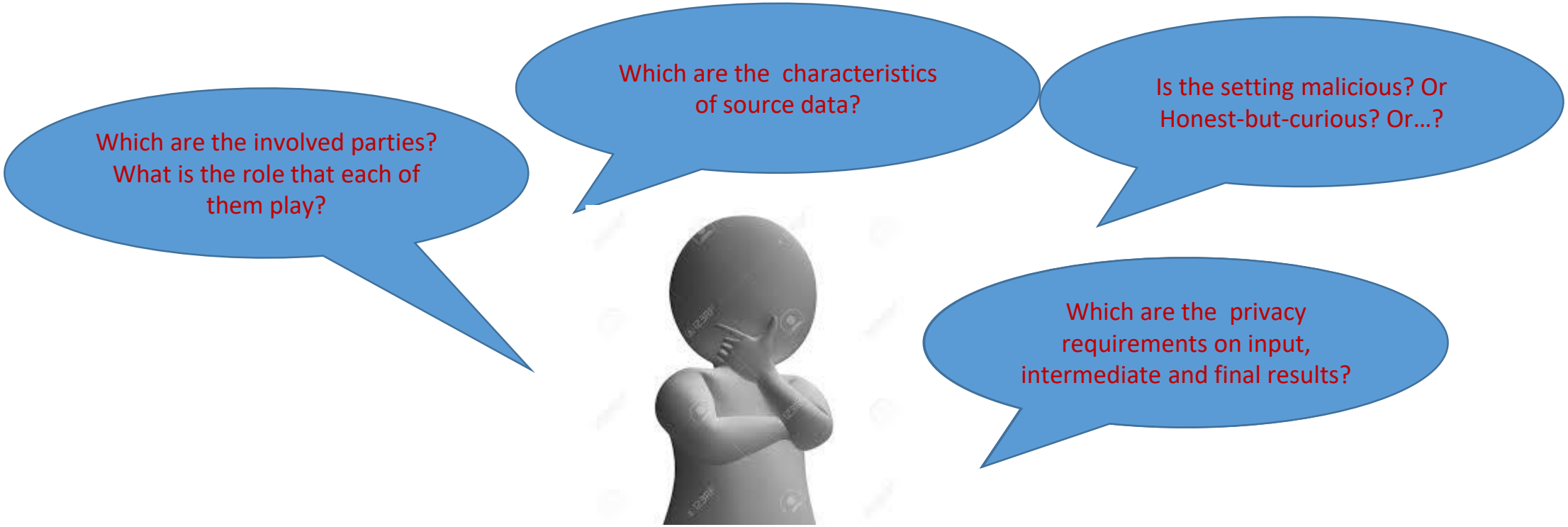
HLG-MOS Project - Input Privacy Preserving Techniques: Work packages

- Input Privacy-Preserving techniques - IPP
- Workpackages:
  - **WP1: To document, generalize and prioritize the use-cases**
  - **WP2: Setup and test the use-cases**
  - **WP3: Lessons learned and guidelines**



Which use cases for input privacy in OS and how to specify them?

- The question is <u>not</u> as easy as it could appear
- Indeed, setting up an IPP scenario requires many details



List of Use Cases

- Use case list from which we started

Use-cases, generalized version
CBS: NSO / Academic researchers joint analyses on sensitive data
CBS: Trusted Smart Surveys
Eurostat: Multi-MNO Statistics
ISTAT: Multi Public Entities -Private Set Intersection with Analytics
StatCan: Private machine learning classification using homomorphic encryption

Logical framework for definition of use cases and scenarios

- Use cases and scenarios are a useful tool for requirement analysis
- We detailed:
  - General Features
  - Solution-specific Features
    - Input parties/Source Data
    - Computing Parties/Statistical analysis
    - Result parties/Statistical products

IPP Project: A Template to specify Input Privacy for OS Use Cases

General features

- **Generic description of objectives of the scenario**
- **General privacy requirements**: these are the requirements that describe how privacy is ensured throughout the whole process, in particular, privacy requirements should be specified for (i) the source data, taking into account identifying and sensitive information that is part of the source data, (ii) statistical analysis, for which it should be specified which information are protected and which are instead potentially available and (iii) statistical products, for which privacy guarantees and possible disclosure risks should be explicitly stated.

General Section: scenario and general privacy requirements

IPP Project: A Template to specify Input Privacy for OS Use Cases

Solution-specific features

- **IPP technique**: the specific choice of the IPP technique.
- **Relationships among (i) Input Organization(s), Output Organization(s) and Computing Entity (ies)**: e.g. Input and Output organizations could be the same and Computing Entity could be a third party.
- **Privacy Threat Types**: The main ones being Linkability, Identifiability, Disclosure of Information.

Solution-specific Section: IPP technique, relationships among parties, privacy threat types

Solution-specific features

Input Parties/Source Data need to be detailed in terms of:

- **Input Organization(s):** i.e. organization(s) providing source data.
- **Type of input parties:** e.g. public or private.
- **Multiplicity of input parties:** i.e. one or more.
- **Source data characteristics:** source data have to be defined in terms of structural characteristics that have an impact on the IPP protocol

The structural characteristics should include:

  - **Data type:** structured or unstructured.
  - **Provision type:** rest or in motion.
  - **Data size:** both horizonal and vertical for structured data.
  - **Metadata:** structured metadata necessary to the IPP protocol.

Solution-specific Section: Input Parties/Source Data Characteristics

IPP Project: A Template to specify Input Privacy for OS Use Cases

Solution-specific features

Computing Parties/Statistical Analysis can be characterized in terms of:

- **Computing Entities:** i.e. parties providing computing capabilities.
- **Type of computing parties**: e.g. *public* or *private.*
- **Multiplicity of computing parties**: i.e. one or more
- **Computing Task characteristics**:

There are at least three relevant dimensions under it is useful to characterize the computing task, namely:

  - **Nature of the task**: additive and multiplicative operations, calculation of a descriptive statistics (mean, variance, median), inference method.
  - **Single vs. multiple datasets:** i.e. those involved in the computations.
  - **Local vs. global** nature of the computation. As an example, a data integration task, like e.g. a record linkage task, has to be intended as a global computation. Instead, there can be local computation like the estimation of the transport mean from accelerometer data on a smart device.

Solution-specific Section: Computing Parties/Statistical Analysis Characteristics

IPP Project: A Template to specify Input Privacy for OS Use Cases

Solution-specific features

Result Parties/Statistical Products can be detailed as follows**:**

- **Output Organization(s)**: i.e. those that benefit from statistical products.
- **Type of result parties**: e.g. *public* or *private.*
- **Multiplicity of output parties**: i.e. one or more.
- **Statistical products characteristics**. Similarly to Source data, statistical products have to be defined in terms of *structural* characteristics that should include:
  - **Data type**: structured or unstructured.
  - **Production type**: rest or in motion.
  - **Data size**: both horizonal and vertical for structured data.
  - **Metadata**: structured metadata necessary to the IPP protocol.

Solution-specific Section: Result Parties/Statistical Products Characteristics

Example: Case Study Istat – Bank of Italy

Istat

Households

A_ID	Number of Children	Age Class
a_{id1}	a_{11}	a_{12}
a_{id2}	a_{21}	a_{22}

Bdl

Mortgage

B_IS	Income Class	Type of Mortgage Payment	Payer Type
b_{id1}	b_{11}	b_{12}	b_{13}
b_{id2}	b_{21}	b_{22}	b_{23}
b_{id3}	b_{31}	b_{32}	b_{33}

Input Parties= Result Parties

Income class aggregates by age class and payer type

F(C)→R

Insolvent payers aggregates by # Children?

$C_{ID}=A_{ID}=B_{ID}$	Number of Children	Age Class	Income Class	Payer type
c_{id1}	a_{11}	a_{12}	b_{21}	b_{23}
c_{id2}	a_{21}	a_{22}	b_{31}	b_{33}

Computing Party

IPP Scenario	Private Set Intersection with Analytics
Scenario's objective	This scenario is Private Set Intersection with Analytics. The parties, named P1 and P2, own databases D1 and D2 respectively. D1 and D2 have a common key, which can be exploited to perform an Exact PSI. The parties wish to enrich their information assets by learning the results of a statistical analysis applied to the intersection of their databases.
General Privacy Requirement	• Only the strictly necessary data are transmitted; • only encrypted data are transmitted; • secure data transmission protocols are used; • the intersection of private databases is obtained by an Exact PSI; • the parties learn only the results of the required statistical analysis (beyond the keys of the records belonging to the intersection); • assuming an HbC environment (i.e. a trustful behavior), it is possible to address the data sharing goal between institutions in a private framework: each party will know either counts with respect to a given set of grouping variables or the actual values of the attributes of records belonging to the other party, with the privacy constraints enforced on identifier fields; • in situations with rarefied distribution of record attributes it could be required the employment of Statistical Disclosure control techniques to assess the risk of reidentification either on the Linker or in the client/server side.
IPP Technique	Multi parti computation with encryption steps*

** E. De Cristofaro and G Tsudik: Practical Private Set Intersection Protocols with linear Computational and Bandwidth Complexity. Proc Financial Cryptography and data Security, 2010.*

IPP Scenario	Private Set Intersection with Analytics
Relationships among (i) Input Organization(s), Output Organization(s) and Computing Entity (ies).	Input Organizations=Output Organizations Computing Entities= Third party
Trust relationships (between Input, Output and Computing parties)	Honest but Curious
Privacy Threat Type	Robust wrt Linkability and Identifiablity
Input Parties/Source Data	
Input Organizations	Istat and Bank of Italy
Type of input parties	Public
Multiplicity of input parties	2
Source data characteristics	Synthetic datasets on Socio-Demographic characteristics (Istat) & Mortgage Contracts (BdI)
Data type	structured
Provision type	rest
Data size	Istat Dataset(10K*4), BI Dataset (15K*3)
Metadata	shared in advance

IPP Scenario	Private Set Intersection with Analytics
Computing Parties/Statistical Analysis	
Computing Entities	NSO
Type of computing parties	Third Party
Multiplicity of computing parties	1
Nature of the task	Exact matching with count by
Single vs. multiple datasets	Two datasets
Local vs. global	Global
Result Parties/Statistical Products	
Output Organizations	Istat and Bank of Italy
Type of result parties	Public
Multiplicity of output parties	2
Statistical Products characteristics	Tables resulting from count_by (e.g. number of family holder that are mortgage holder by profession)
Data type	structured
Provision type	rest
Data size	(less than 10K,3)

Conclusions and Next Steps

- Specifying such use cases is not a trivial task, hence a template for supporting their definition is a valid support
- The IPP project proposed a template for specifying input privacy for OS use cases as a first step towards the definition of a logical framework for dealing with Input Privacy for OS
- Next steps of the IPP project to evolve the framework:
  - Guidelines to apply Input Privacy techniques in OS scenarios

Thanks!

MONICA SCANNAPIECO
monica.scannapieco@istat.it