Data Protection and Privacy

University of Genoa

Lesson 6: Data Generation

Gaspare Ferraro <ferraro@gaspa.re>

Data generation I

- There are areas where anonymization techniques are not secure enough to protect against threats
- The alternative, in some case, is to generate synthetic data
- Synthetic data is information that's artificially generated rather than produced by real-world events
- Scientific researchers in fields such as physics and chemistry have used artificial data to supplement the unavailability of experimental data
- It is observed that although easy to create, artificial data can lead to results that are either significantly different or opposite in certain cases

Data generation II

- The necessity of synthetic data arises due to high volume, high sensitivity of data, no historical data availability, and bridging incomplete data
- There are tools used today to generate synthetic data
- Synthetic data generation is not limited to just relational data, there are other forms of data such as graph data, time series data, spatial-temporal data, and longitudinal data, which may require synthetic data for reconstruction, data bridging, or testing
- We will analyze several techniques and aspects to consider when dealing with synthetic data generation

Privacy and Utility in Synthetic Data

- The first step in creation of synthetic data is to clearly classify the metadata
- After classification, there are a few steps that need to be followed to build a model for synthetic data generation
- The model consists of rules that are created using:
  - Metadata information
  - Referential integrity constraints
  - Dependent and independent variables
  - Correlated fields data
  - Domain context
  - Application scenarios
- Synthetic data are generated according to rules. The model represents the application context in the form of instructions that help to produce useful data.

Dealing with Explicit Identifier generation

- Creating EIs is the least complicated part as very little is analyzed or tested using them. However, this does not reduce the importance of their privacy preservation. EIs are relatively unimportant, regardless of the purpose data are used for.
- In data mining, specific names, social security numbers (SSNs), and phone numbers are not important. It is the collective profile of record owners that are important, which cannot be obtained from EIs
- Therefore, privacy is more important to preserve in the case of EIs than utility
- Synthetic data generation techniques are very specific to the field type and format
- EIs could be free text fields such as names or specifically formatted ones such as SSNs

in sostanza devi essere consistente: se crei un finto dato su un italiano devi usare un numero italiano quindi con certe strutture stessa cosa per il documento ecc. ovviamente i dati non devono avere senso solo fra di loro ma anche in generale es l'email ha un certa struttura che deve essere seguita

Explicit Identifier Generation Techniques

SDG Technique	Brief Explanation
Substitution	Prepopulated sets of data are created. For example, first name, last name, and middle name that are directly substituted for original data. Substitution is difficult to implement when consistency is a requirement due to the randomness involved in picking the replacement.
Credit card, social security number, aadhar number	Format is preserved while replacing original digits and characters with randomly generated ones. The randomness needs to be carefully introduced wherever the meaning of data is not impacted. A PAN number in India has meaning for some characters in it. For example, the fourth character "P" in AAZPE3479P means the PAN belongs to an individual, whereas a 'C' in the same place means it belongs to a company.
E-mail address	Based on standards being followed, either an e-mail address is generated for entire record set uniquely or a common e-mail address is assigned to each row.
Mobile phone numbers	Most often, mobile numbers are 10-digit numbers that can be generated using a random number generator.
Flat value	A single value is assigned to all rows within a set.

Dealing with Quasi-Identifiers generation

- Generate EIs requires only conforming to the syntax of the field
- Generate QIs requires also to preserve the semantics as well
- It is important to collect metadata to understand the syntax and statistical distribution of the original data, that is, boundary values, mean, and variances of each numerical attribute
- Categorical attributes require a higher level of analysis, as their meaning or spread of values may be important

vanno anche seguite regole statistiche per rendere i dati realistici. Es se metto l'età non posso avere tutti che hanno 100 anni andrà seguita la reale distribuzione di età del mondo

oppure la distribuzione di abitanti, non abitano tutti su un isola del pacifico

Privacy of Quasi-Identifiers generation

- QIs are data that help identify a record owner when combined with background knowledge or external data sources
- While generating synthetic data, an important aspect is to maintain some distance between original data and synthetic data
- For QIs to be highly private, all values need to be distorted to a certain extent that they do not compromise any individual identities as a result of external linkage
- The extent of distortion can be measured using appropriate distance functions

Utility of generated Quasi-Identifier

- Synthetic data for testing need to capture all scenarios required to invoke various program flows
- Similarly for data mining, analysis and the relationship between QIs and SD are important to create good data
- The utility of QIs is linked to the preservation of the correlation with SD fields
- Correlation may also exist between groups, as in a set of independent variables and a set of dependent variables
- As the number of these variables increases, so does the complexity of generating synthetic data
- Correlations among QI and SD fields are the most common

Quasi-Identifiers Data Generation Techniques

SDG Technique	Brief Explanation
Randomization	A random value is generated to replace the original value. Based on the data format and range of original data, a number is randomly picked. This technique is applicable to date fields specifically.
Geographical area	Addresses and zip codes can be generated to be valid zip codes, or if the data are too clustered (like the voters list of a particular city), geographical area can be substituted against a zip code. Of course, the prerequisite is that altering the format of data is permissible.
Generalization	Generalizing a set of nominal categorical data values like making master's degree or PhD into postgraduate.

Sensitive Data Generation

- Sensitive data form the most important ingredient of a data set.
- In data mining, facts are always attributed to sensitive data while profiling is attributed to Quasi-Identifiers.
- Numerical SD attributes are easier to generate
- While additional rules needs to be developed to categorical SD

Privacy of Sensitive Data generation

- Synthetically generated SD are similar to that discussed in QIs
- Distribution of univariate and multivariate data is a good guideline to synthetically create data
- Single column data can be easily created, but preserving their association with correlated data is important
- It is always desirable to have synthetic data created as close to original values as possible
- However, from a privacy preservation perspective, the closer the SD to original values, the greater the privacy concerns are