Detection of Data Leaks in Collaborative Data Driven Research

by Peter Kieseberg, Edgar Weippl, SBA Research, and Sebastian Schrittwieser, TARGET

Collaborative data driven research and development is one of the big issues in BigData-Analysis. Still, several obstacles regarding the preservation of ownership over the shared data need to be overcome to enable the techniques to unfold their full potential. This especially concerns the detection of partners that leak valuable data to the outside world.

In the current age of big data, machine learning and semantic systems are playing increasingly important roles in data analysis. In the medical sector in particular, large volumes of data are stored in well-protected databases at various institutions, such as hospitals and research facilities. But more traditional industries, such as production environments and factories, have also increased the incorporation of sensor data for optimization purposes or in order to make new features and services available. The industry 4.0 paradigm is often seen as a trendsetter for the next decade, with different speeds of adoption by different industries.

Scientific analysis of data sets could often be improved tremendously by cooperating with other institutions like research facilities or suppliers, which either have access to additional data (e.g. complementary information on the production process of raw material) or are researching new analytical methods or services. Especially when developing algorithms in the area of machine learning, the possibility of tailoring the algorithm to the specific problems, including investigating the best definition for the parameters, can have a vast impact on the overall performance. Thus, many industrial development and research programs could profit from exchanging relevant data between the participants.

While data protection is a well-researched area, see for example [1], the problem remains that even when using proper anonymization, the leakage of data sets can lead to severe problems. The reasons for this range from financial concerns, e.g. when the data sets in question are offered for sale, to strategic reasons, e.g. the anonymized data provides information about the interests and/or capabilities of the institution or industry, to the desire to keep control over the distribution of the data. Additionally, the threat of collusion attacks that can be used to subvert anonymization arises when differently anonymized data sets are provided to multiple data analysts. Fingerprinting techniques can be used in order to mitigate the risk of unwanted data leakage by making it possible to attribute each piece of distributed data to the respective data recipient.

While some approaches exist for fingerprinting structured data as is typically stored in tables, these usually rely on the availability of a substantial portion of the leaked data to identify the leak. In the DEXHELPP-project [L1], we thus conduct research on a combined approach for anonymization and fingerprinting in a single step, as both needs are apparent in many data driven environments: The data set for each recipient is anonymized slightly differently using k-anonymity based on generalization and providing roughly the same remaining data quality, thus resulting in data sets with different anonymization strategies but close to equal value (see Figure 1).

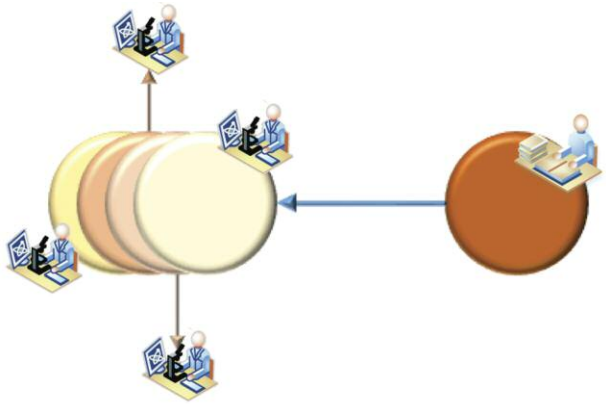

Figure 1: Data owner provides different versions to recipients.

Since generalization achieves k-anonymity by applying generalization to each individual record in the data set, only a single record is needed to detect and identify the leaking data recipient (see Figure 2):

(A) A data record is encountered in the wild.
(B) The generalization strategies for the field attributes are identified and the resulting identification pattern is generated.
(C) The identification pattern is matched against a list holding all data recipients and the patterns of their respectively received data sets, thus allowing identification of the leaking recipient.

One of the main problems of this approach lies in the danger of colluding attackers and resulting inference attacks. Thus, the generalization patterns must be selected in such a way that any arbitrary combination of these data sets neither results in a reduction of the anonymization level k, nor in the ability to hide the identities of the colluding attackers (see [2]). Furthermore, the theoretical approach based on utilizing generalization and k-anonymity will be extended to incorpo-

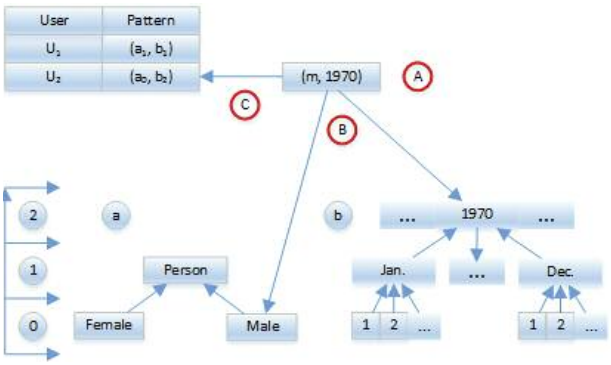

Figure 2: Detection of the data leak.

rate more advanced concepts for anonymization, as well as more diverse mechanisms for achieving anonymity. The main reason for this research step lies in the various other prerequisites for data anonymization, especially in the medical sector, that are not fulfilled by basic k-anonymity, e.g. considering dependencies between data records or statistical attacks.

During the DEXHELPP-project, the devised algorithms will not only be tested with synthetic data, but also with real medical records. This is especially important, as the actual performance of this fingerprinting approach largely depends on the ability to cluster the data into equivalency classes that form a partition of the whole data set. This again relies very much on the actual distribution of the data inside the data set, especially when refraining from removing outliers. Principal performance identifiers are not only related to KPIs that are typically identified with performance, but mainly with attributes like the number of possible fingerprints in a given data set, the distance between two fingerprinted sets, as well as the achievable remaining quality.

In conclusion, our work on fingerprinting data sets in the DEXHELPP project will result in practical fingerprints for collaborative data driven research and development environments. The resulting best fingerprinting strategy will be implemented inside the DEXHELPP test server environment, which was designed to securely perform collaborative medical research on data provided by the Austrian healthcare providers whilst preserving privacy.

Link:
[L1] http://dexhelpp.at/?q=en

References:
[1] B. Fung, et al.: "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys (CSUR) 42(4), 14 (2010).
[2] P. Kieseberg, et al.: "An algorithm for collusion-resistant anonymization and Fingerprinting of sensitive microdata", Electronic Markets - The International Journal on Networked Business, 2014.

Please contact:
Peter Kieseberg, SBA Research, Vienna, Austria
pkieseberg@sba-research.org

HOBBIT: Holistic Benchmarking of Big Linked Data

by AxelCyrille Ngonga Ngomo, InfAI, Alejandra García Rojas, ONTOS, and Irini Fundulaki, ICS-FORTH

The HOBBIT project aims at abolishing the barriers in the adoption and deployment of Big Linked Data. To this end, HOBBIT will provide open benchmarking reports that allow to assess the fitness of existing solutions for their purposes. These benchmarks will be based on data that reflects reality and measures industry relevant Key Performance Indicators (KPIs) with comparable results using standardized hardware.

Linked Data has grown rapidly over the last ten years [L1]. Organizations are increasingly interested in using solutions based on Linked Data. However, choosing the right solution for their needs remains a difficult task. HOBBIT's rationale is to support organizations that aim to use Linked Data technologies at all scales (including Big Data) in the choice of appropriate solutions. To this end, HOBBIT [L2] will provide benchmarks for all the industry relevant phases of the Linked Data lifecycle [1] according to the approach shown in Figure 1.

In particular, the H2020 Hobbit EU project will create benchmarks for the following stages:

1. Generation and Acquisition: Benchmarks pertaining to the transformation of unstructured, semi-structured and structured data into RDF.
2. Analysis and Processing: Benchmarks pertaining to the use of Linked Data to perform complex tasks such as supervised machine learning.
3. Storage and Curation: Benchmarks pertaining to the storage, versioning and querying of RDF data stored in corresponding solutions.
4. Visualisation and Services: Application centric benchmarks pertaining to queries used by software solutions which rely on large amounts of Linked Data.

The HOBBIT project will provide innovative benchmarks based on the following premises:

- Realistic benchmarks: Benchmarks are commonly generated with synthetic data that reflect a single and specific domain. HOBBIT aims at creating mimicking algorithms to generate synthetic data from different domains.
- Universal benchmarking platform: We will develop a generic platform that will be able to execute large scale benchmarks across the Linked Data lifecycle. The platform will provide reference implementations of the KPIs as well as dereferenceable results and automatic feedback to tools developers..
- Industry relevant Key Performance Indicators (KPIs): In addition to the classical KPIs developed over the last decades, HOBBIT will collect relevant KPIs from industry to make the assessment of technologies based on the industrial needs possible.