# Privacy Preserving OLAP Data Cubes

**Alfredo Cuzzocrea**
*ICAR-CNR and University of Calabria, Italy*

## INTRODUCTION

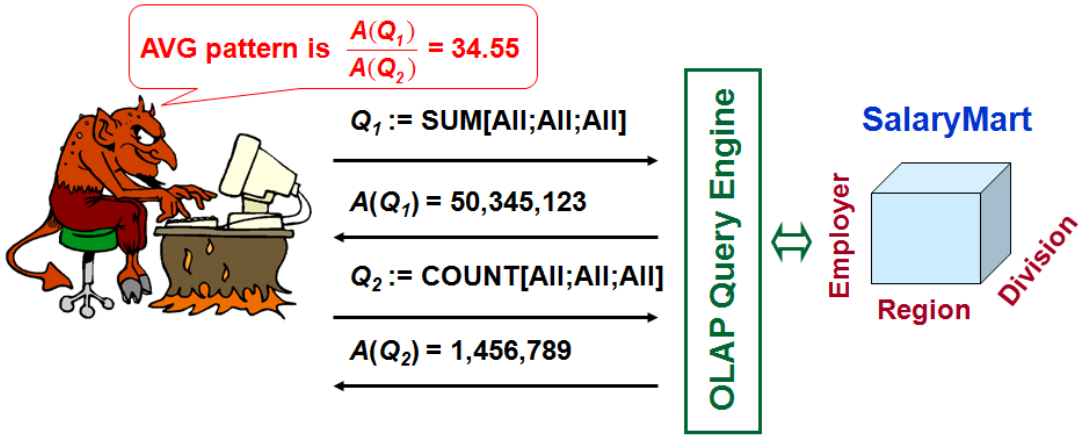
It has been demonstrated (Sweeney, 2002) that malicious users can infer *sensitive knowledge* from online corporate databases and data cubes that do not adopt effective privacy preserving countermeasures. From this breaking evidence, a plethora of *Privacy Preserving Data Mining* (PPDM) (Agrawal & Srikant, 2000) techniques has been proposed during the last years. Each of these techniques focuses on supporting the privacy preservation of a *specialized* KDD/DM task such as *frequent item set mining*, *clustering* etc. *Privacy Preserving OLAP* (PPOLAP) (Agrawal, et al., 2005) is a specific PPDM technique dealing with the privacy preservation of data cubes (Gray et al., 1997). Data cubes play a leading role in *Data Warehousing* (DW) and *Business Intelligence* (BI) systems, as, on the basis of a *multidimensional* and a *multi-resolution* vision of data, data cubes make available to OLAP users/applications SQL aggregations (e.g., SUM, COUNT, AVG etc) computed over very large amounts of data stored in data sources (e.g., relational databases). These aggregations enable OLAP users/applications to easily extract *summarized knowledge* from the underlying massive data sources, with performance infeasible for traditional OLTP processes. Unfortunately, as highlighted by recent studies (Pernul & Priebe, 2000; Wang et al., 2004a; Wang et al., 2004b; Agrawal et al., 2005; Hua et al., 2005; Sung et al., 2006), the privacy risk heavily affects online-published data cubes. By accessing and querying data cubes, malicious users can infer OLAP aggregations computed over sensitive ranges of multidimensional data that, due to privacy reasons, are hidden to unauthorized users. Specifically, since OLAP deals with aggregate data and summarized knowledge, malicious users are usually interested in inferring what we define as *aggregate patterns* of multidimensional data, rather than *individual information of data cells stored in data cubes* (e.g., (Sung et al., 2006)) *or tuples stored in relational databases* (e.g., (Sweeney, 2002; Machanavajjhala et al., 2007)). Given a multidimensional range $R$ of a data cube $A$, an aggregate pattern over $R$ is defined as an aggregate value extracted from $R$ that is able of providing a "description" of data stored in $R$.

Consider the following example case study, which is depicted in Figure 1. Here, a three-dimensional corporate data cube storing salary data, called *SalaryMart*, which is characterized by the dimensions set $D = \{Employer, Division, Region\}$ and the measure $M = \{Income\}$, is accessed by a malicious user via a conventional OLAP query engine. By exploiting the knowledge about data cube metadata (such as dimensions, along which their definition set and cardinality, cardinality of the data cube, and so forth) and query metadata (such as dimensions, selectivity, and so forth), and thanks to the rich availability of OLAP operators and tools (e.g., (Chaudhuri & Dayal, 1997)), malicious users can infer (yet-approximate) aggregate patterns, by realizing what we call as *simple attacks* to OLAP data cubes. Figure 1 shows an example of such attacks, where the malicious user is able to retrieve the (yet-approximate) value of the AVG pattern of the data cube *SalaryMart* by means of simple *linear-interpolation-based query answering methods over data cubes* (e.g., (Cuzzocrea, 2006; Cuzzocrea & Wang, 2007)).

This attack model can evolve towards more problematic *complex attacks* to OLAP data cubes, via meaningfully exploiting further knowledge that, due to pre-fixed business processes of the

DOI: 10.4018/978-1-4666-5202-6.ch169

*Figure 1. An example simple attack to an OLAP data cube*




target organization, is disclosed to the outside. A kind of such knowledge is modeled thorough a set of queries that are made available to external users. We name these queries as *a-priori-known queries*. Starting from the knowledge about the answers to these queries, malicious users can infer (yet-approximate) answers to so-called *inferred queries*, which, due to privacy reasons, are hidden to unauthorized users. For instance, Figure 2 shows an example complex attack to the data cube *SalaryMart*, where the answer to the (inferred) query $Q_3$ is inferred from the answers to the (a-priori-known) queries $Q_1$ and $Q_2$, again based on linear interpolation tools over data cubes (e.g., (Cuzzocrea, 2006; Cuzzocrea & Wang, 2007)).

Malicious OLAP scenarios like the ones described above get even worse when data cubes are processed in a distributed setting, in the context of so-called *Privacy Preserving Distributed Data Mining* problem (Clifton et al., 2002) where the main goal is to efficiently support Data Mining activities (e.g., *classification*, *clustering*, *frequent item set mining*, *OLAP*, *OLAM*) across multiple distributed databases while ensuring that (*i*) no participant can access sensitive data stored in databases of other participants, and (*ii*) no participant can infer sensitive knowledge other than the knowledge obtained by the target Data Mining activity.

*Figure 2. An example complex attack to an OLAP data cube*