Chapter 8.3
Advances in Privacy Preserving Record Linkage

Alexandros Karakasidis
University of Thessaly, Greece

Vassilios S. Verykios
University of Thessaly, Greece

ABSTRACT

The contemporary era is characterized by a high degree of involvement of computers harvesting data in various aspects of everyday life. Merging these data would provide numerous benefits in various fields, such as in the field of medical research, where new patterns for diseases could be established. However, this is not a trivial task, since among other reasons, separate data holders maintain data corresponding to the same real world entities without necessarily maintaining common and unique linkage identifiers. Additionally, these data may contain errors, rendering the integration process very difficult. The aforementioned reasons encountered during merging data from heterogeneous sources comprise some important aspects of the classical linkage problem.

However, even though many solutions have been proposed towards addressing this problem, a new side effect rises regarding the privacy of the data which usually has to be protected during linkage. Sensitive information such as names, addresses, and illnesses, especially in cases of medical data, should not be revealed without further evidence to any participant of the merging procedure. This raises the need of creating new techniques for linking data while, at the same time, the privacy of the subjects described by these data is preserved. This need led to the evolvement of a new research area called privacy preserving record linkage. This chapter will attempt to present the state of the art of the methods proposed to address the privacy preserving record linkage problem and provide a taxonomy of these techniques based on their core characteristics.

DOI: 10.4018/978-1-61350-323-2.ch8.3

1 INTRODUCTION

Nowadays, both public and private organizations maintain databases consisting of information for every one of us. Very often these organizations need to integrate these data. The reasons for such an action may vary from scientific purposes to performing market surveys. In any case, however, the privacy of the individuals described by these data should not be compromised.

An important field of application of data integration concerns the sector of public health and safety. Gathering information for medical research would have as a result to facilitate research towards establishing patterns for diseases. Moreover, being able to combine all this independent information could lead to the creation of a public safety early warning system as Clifton et al. (Clifton et al., 2004) and Bhowmick et al. (Bhowmick, et.al, 2006) describe. The ability of building such a system would also be useful as a component for other systems for conducting surveys either for commercial or scientific reasons.

Very often, companies need to merge their data in order to redefine their marketing policies. In such a situation, the matching parties may not wish to reveal their databases to each other for competitive reasons. Additionally, customer information should not be freely exchanged since this is considered as a privacy violation. All these aspects comprise part of a hot problem known as privacy preserving record linkage.

It would be useful to distinguish the difference between private record linkage and private data mining. While in private record linkage the aim is to obfuscate data maintaining at the same time their usability in order to perform data integration, in private data mining the aim is to preserve privacy of personal information during the data mining process (Verykios et al., 2004). In other words, privacy preserving record linkage is a step prior to privacy preserving data mining.

At this point we would like to recommend for the novice reader for this domain, fundamental works concerning both the classical and the private record linkage field. More specifically, considering the classical record linkage problem we suggest the works of Herzog et al. (Herzog et al., 2007) and Elmagarmid et al. (Elmagarmid et al., 2007). We also suggest the work of Christen (Christen, 2008) which contains details regarding Febrl, a record linkage toolbox. Concerning the private record linkage field, the interested reader should consider the works of Clifton et al. (Clifton et al., 2004), Du et al. (Du et al., 2000) and Kantarcioglou et al. (Kantarcioglu et al., 2008).

2 BACKGROUND

To introduce the reader to the problem of privacy preserving record linkage, we will provide in this section some elementary materials regarding the issues which need to be addressed for its viable solution.

2.1 Preliminaries

Let us take a first look at the challenges involved in addressing the privacy preserving record linkage problem. First of all, identically to the classical record linkage problem, the databases that are going to be integrated do not share common primary keys. Therefore specific steps regarding linkage keys to be selected have to be employed as described in (Trepetin, 2008).

Record linkage is the process of identifying the same real world entity in two or more separate databases. The first problem rising in such a situation is that these databases do not have the same primary keys, a fact that forces us to examine other ways for joining. First of all, unique linkage identifiers may be used such as the SSN. Since there might be errors or formatting discrepancies between the matching databases, these identifiers could be examined for similarity used some predefined metric, able to assess if they refer to the same real world entities.