Anonymity and Pseudonymity in Data-Driven Science

Heidelinde Hobel
SBA Research, Austria

Sebastian Schrittwieser
St. Poelten University of Applied Sciences, Austria

Peter Kieseberg
SBA Research, Austria

Edgar Weippl
SBA Research, Austria

INTRODUCTION

Rapidly emerging technological improvements, e.g., in the area of data storage, computation performance, cloud computing and collaborative work supporting technologies, as well as new methods in statistical analysis, facilitate a new field of science we refer to as data-driven science. The general objectives of this specific area are to reason based upon empirical findings and evidence for assumptions by analyzing huge amounts of data that are nowadays familiar by the term "Big Data". Big Data can be derived from several sources, e.g., the Web, research partners, volunteers or enterprises. In general, noncommercial research is meant to be released to the public, but often only the actual results of the analyzed data are presented and/or the data included in condensed or abstracted form in a way that it is impossible for the reader to repeat the evaluation for validation or personal learning effect. The peer review procedure should ensure quality of research, but without access to the underlying research data it is virtually impossible to perform this fundamental element of proper validation for the reviewer in the field of data-driven science. Especially in academic research the phrase "publish or perish" describes the pressure of scholars to publish new publications frequently, otherwise they will not be recognized in the academic field or by funding institutions. That leads to the assumption that without publishing the research data, at least for the review process, we leave open a hole for fraud and poor research. However, even if a policy would be enforced that the underlying research data has to be available for the public or maybe limited to reviewers, it would imply that sensitive or data subject to privacy polices is still made public, even though maybe only to a limited group. Once access to data has been provided, it cannot be withdrawn and the possibility is higher that the data can be exploited for malicious purpose. Likewise, in cooperative research the data has to be published at a need-to-know basis, adhering to privacy policies. Therefore, we provide an overview on the latest data anonymization and pseudonymization techniques that should prevent data disclosure and inference attacks.

BACKGROUND

According to recent publications, science is becoming more data-driven (Bonneau, 2012; Chia, Yamamoto, & Asokan, 2012; Dey, Zubin, & Ross, 2012; Siersdorfer, Chelaru, Nejdl, & Pedro, 2010; West & Leskovec, 2012; Zang & Bolot, 2011). The term "Big Data" has originally emerged

DOI: 10.4018/978-1-4666-5202-6.ch013

from the IT-sector, where large data samples had to be analyzed. In many publications, large data sets are used to evaluate proposed prototypes or algorithms, especially concerning practical applicability and with regards to performance issues. Furthermore, they also serve as underlying research foundation for new empirical findings that can be derived from analyzing the data for general trends and characteristics. However, we can learn from privacy-preserving health data publishing where sensitive data from patients has to be protected from leaking into the public (Sweeney, 2002a; Sweeney, 2002b). From this we learned that not only direct identifiers, like the social security numbers, may contribute to the threat of a privacy breach, but also quasi-identifiers (QI), e.g., the triple ZIP, birth date and gender, could lead to a possible identification of a person so that private data like diseases about patients could be inferred about identified patients for malicious purpose (Sweeney, 2002a; Sweeney, 2002b). But not only in health care huge amounts of data could lead to new findings or evidence for assumptions. For instance, Dey et al. (2012) have analyzed 1.400.000 Facebook account settings in order to infer privacy trends for several personal attributes. Although public accounts were used for their research methods, their results combined with their measured and recorded data are highly sensitive and should not be published without appropriate anonymizations or pseudonymization techniques. The effort of building relationships between the sensitive published data and data that is public or easily accessible for attackers is denoted as data linkage (Fung, Wang, Chen, & Yu, 2010). Altogether, as we depend on the data disclosure of volunteers, we are responsible for preserving data privacy. This includes ensuring the unlinkability of sensitive data such that the data records can be published to facilitate validation of research, collaboration between several science groups and for the personal learning effect by enabling the reader to repeat the proposed scientific experiment.

PRIVACY-PRESERVING DATA PUBLISHING IN DATA-DRIVEN SCIENCE

A

Let us assume that research data comprises a set of attributes that could be structured in identifiers, QIs, sensitive and non-sensitive attributes as well as inferred knowledge that had been gained in the research approach. For instance, considering the above-mentioned Facebook example, we assume that each data record comprising all required data of an account has been classified according to predefined privacy categories. Thus, it is obvious that the links to the accounts are classified as identifiers which have to be removed before publishing. Additionally, the concept of QIs that was initially proposed by Sweeney (2002a) could lead to a possible identification of an account by using a large set of the published attributes and simply matching these keywords with words on public Facebook accounts. This specific attack is called inference attack or linkage where background knowledge, a-priori-knowledge or public data is used for entailing the identity of the record owner. In the above-mentioned example, an identified account would be a privacy breach where the attacker eventually acquired a scientifically approved privacy classification of the published accounts.

In the following we distinguish between four inference attacks - record, attribute, and table linkage as well as probabilistic attacks (Fung et al., 2010). In a record linkage attack, the QI is linked directly with additional knowledge such as the above-mentioned keyword matching example of Facebook. Attribute linkage relates to the threat between a possible correlation of QIs and inferred knowledge. For instance, if one has the background knowledge that a particular person is within a certain equivalence group, the person's sensitive attribute can easily be identified. Table linkage relates to attacks where the presence or the absence of the record owner is entailed whereas probabilistic attacks refer to the threat of a change