Chapter 10

Privacy Inference Disclosure Control with Access-Unrestricted Data Anonymity[1]

Zude Li
The University of Western Ontario, Canada

ABSTRACT

This chapter introduces a formal study on access-unrestricted data anonymity. It includes four aspects: (1) it analyzes the impacts of anonymity on data usability; (2) it quantitatively measures privacy disclosure risks in practical environment; (3) it discusses the potential factors leading to privacy disclosure; and (4) it proposes the improved anonymity solutions within typical k-anonymity models, which can effectively prevent privacy disclosure that is related with the published data properties, anonymity principles, and anonymization rules. The experiments have found these potential privacy inference violations and shown the enhanced privacy-preserving effect of the new anti-inference policies to access-unrestricted data publication.

INTRODUCTION

The term "*privacy*" is generally used as "the right to select what personal information about me is known to what people" (Westin, 1976). In a technical view, data privacy protection is to manage individual data in a privacy aware way, including determination of *when*, *how* and *to what extent* such data can be communicated to others with *what special features*. Privacy is becoming a hot concern for both individuals and organizations.

In the past thirty years, a lot of research studies have been contributed to data privacy. Roughly, these studies can be classified into two types: *access-restricted* and *access-unrestricted*. The *access-restricted* type focuses on how to protect privacy through restricting user access to the private data: A user can access the data only if he/she satisfies the access constraints predefined on the data. The *access-unrestricted* type concerns on data publication, making them oriented to most users and at the same time avoiding possibly compromised privacy disclosure. In the access-restricted type studies, researchers usually

DOI: 10.4018/978-1-61692-000-5.ch010

extend access-based control models for privacy enhancing (Antón et al., 2003; Crook et al., 2005). For instances, *Purpose-Based Access Control* (PBAC) (Byun et al., 2005) and *Hippocratic Database* (HDB) (Agrawal et al., 2002) are two *access control* based privacy techniques. In these models, *direct privacy disclosure* (i.e., access some unauthorized data or with illegal purposes) can be efficiently prevented, while *indirect privacy disclosure* (i.e., infer unauthorized or purpose-illegal private data based on accessed data aggregation) is still a challenging problem. Some *inference control* techniques (Staddon, 2003) are available for handling this issue.

Access-unrestricted data privacy techniques are widely required in practice, such as individual data dissemination, voting data proclamation, health-care data publication, etc. Data anonymity is generally the privacy solution used for this type of applications. The key issue for enhanced access-unrestricted privacy protection is to quantitatively measure the risk of privacy disclosure and information loss which are coupled by data anonymity. One of recently proposed privacy techniques for handling access-unrestricted data privacy is *k-anonymity* (Sweeney, 1997). In nature, it is a data processing model towards maintaining data usability as well as avoiding privacy disclosure during unrestricted data access. No matter of its implementation complexity in practice, privacy disclosure on k-anonymized access-unrestricted dataset is still existed (Machanavajjhala et al., 2006; Li et al., 2006a).

This chapter aims to a formal research on access-unrestricted privacy protection, mainly focusing on the data anonymization process and the privacy inference risk analysis with anonymized data. As the best we know, there are no literatures related to privacy inference control study for access-unrestricted privacy applications. Most anonymity risk detection techniques focus on the anonymization process on only the resulting dataset but not any other external information, which, subsequently, cannot guarantee the survivability of real application systems. This chapter, based on our early work (Ye et al., 2007), illustrates an approach to measuring privacy inference risks on the k-anonymized dataset in a quantitatively manner. Through the data anonymity analysis, we discover four main factors that may incur various privacy violations. Further, we propose two effective *anti-inference* privacy policies for access-unrestricted data publication. In this chapter, k-anonymity is used as a scenario model to perform data anonymity. In the experiments studied, we have proven the existence of potential privacy inference violations and the efficiency of our anti-inference policies.

The rest of this chapter is organized as follows. Section 2 discusses the related work. Section 3 defines some useful concepts and notations for the data anonymity analysis; Section 4 discusses the anonymity principles and anonymization rules that are explicitly or implicitly used during data anonymity; Section 5 illustrates the formal model of access-unrestricted data anonymity; Section 6 analyzes the data anonymization process and the corresponding privacy inference attacks with the k-anonymity model, in which experimental findings are shown for discovering and removing potential various privacy violations and convincing the advantages of our proposed anti-inference policies over the existing ones; Finally, Section 7 gives a short conclusion for the chapter.

RELATED WORK

In access-unrestricted microdata publication applications, all privacy attacks have the common feature: to find the exact or enough-precise mapping between individual identities (*UI*) and sensitive values (*SA*) (Sweeney, 1997; Meyerson & Williams, 2004; Staddon, 2003; Rastogi et al., 2007; Mcguckin & Nguyen, 1990). From the perspective of privacy protection, data publisher should prevent users from them through some special data processing techniques.