IGI PUBLISHING ITJ 4751

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, International Journal of Information Security and Privacy, Volume 2, Issue 4
edited by **Hamid R. Nemati © 2008, IGI Global**

On Access-Unrestricted Data Anonymity and Privacy Inference Disclosure Control

Zude Li, University of Western Ontario, Canada

Xiaojun Ye, Tsinghua University, China

ABSTRACT

This article introduces a formal study on access-unrestricted data anonymity. It includes four aspects: (1) analyzes the impacts of anonymity on data usability; (2) quantitatively measures privacy disclosure risks in practical environment; (3) discusses the factors resulting in privacy disclosure; and (4) proposes the improved anonymity solutions within typical k-anonymity model, which can effectively prevent privacy disclosure that is related with the published data properties, anonymity principles, and anonymization rules. With the experiments, the authors have proven the existence of these potential privacy inference violations as well as the enhanced privacy effect by the new anti-inference policies for access-unrestricted data publication.

Keywords: *Anonymity; Data Privacy; Data Publication; K-Anonymity; Inference Control*

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, 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 extend access-based control models for privacy enhancing (Antón, Bochini,

& He, 2003; Crook, Ince, & Nuseibeh, 2005). For instances, *Purpose-Based Access Control* (PBAC) (Byun, Bertino, & Li, 2005) and *Hippocratic Database* (HDB) (Agrawal, Kiernan, Srikant, & Xu, 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, and so forth. 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-unstricted data parivacy 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 (Li, Zhan, & Ye, 2006a; Machanavajjhala, Gehrke, & Kifer, 2006).

This article 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 article, based on our early work (Ye, Li, Zhan, & Li, 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 variouse privacy violations. Further, we propose two effective *anti-inference* privacy policies for access-unrestricted data publication. In this article, 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 article follows. The Concepts and Notations section defines some useful concepts and notions for the formal data anonymity analysis; the Anonymity Principles and Rules section discusses the anonymity principles and anonymization rules that are explicitly or implicitly used during data anonymity; the Data Anonymity Analysis section illustrates the formal model of access-unrestricted data anonymity; the Data Anonymity vs. Privacy Inference section analyzes the data anonymization process and the corresponding privacy inference attacks with k-anonymity model, in which experiments are depicted for discovering and removing potential various privacy violations and convincing the advantages of our proposed anti-inference policies over the existing ones; the Related Work section discusses the related work. Finally, the Conclusion section gives a short conclusion for the whole article.

CONCEPTS AND NOTATIONS

Suppose individual data are collected into a relational table in database, where each record maps to an individual in reality. For such a special table, the possible data attributes can be divided into three classes (similar to (Domingo-Ferrer, Oganian, & Torra, 2002)): (1) *Unique identifier* attributes (*UI*), where each value refers to an individual in a scope; (2) *Quasi-identifier* attributes (*QI*), where each value can be joined with other values to re-identify an individual in the scope with high precision; and (3) *Sensitive* attributes (*SA*), where each value should not be