Chapter 5
Privacy Hash Table

Xiaoxun Sun
Australian Council for Educational Research, Australia

Min Li
University of Southern Queensland, Australia

ABSTRACT

A number of organizations publish microdata for purposes such as public health and demographic research. Although attributes of microdata that clearly identify individuals, such as name, are generally removed, these databases can sometimes be joined with other public databases on attributes such as Zip code, Gender, and Age to re-identify individuals who were supposed to remain anonymous. These linking attacks are made easier by the availability of other complementary databases over the Internet. K-anonymity is a technique that prevents linking attacks by generalizing or suppressing portions of the released microdata so that no individual can be uniquely distinguished from a group of size k. In this chapter, we investigate a practical full-domain generalization model of k-anonymity and examine the issue of computing minimal k-anonymous solution. We introduce the hash-based technique previously used in mining associate rules and present an efficient and effective privacy hash table structure to derive the minimal solution. The experimental results show the proposed hash-based technique is highly efficient compared with the binary search method.

INTRODUCTION

Several microdata disclosure protection techniques have been developed in the context of statistical database, such as scrambling and swapping values and adding noise to the data while maintaining an overall statistical integrity of the result (Adam et al. 1989). However, many applications require release and explicit management of microdata while maintaining truthful information within each tuple. This data quality requirement makes inappropriate those techniques that disturb data and therefore, although preserving statistical properties, compromise the correctness of the single pieces of information. Among the techniques proposed for providing anonymity in the

DOI: 10.4018/978-1-61350-501-4.ch005

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

release of microdata, we focus on two techniques in particular: generalization and suppression (in the Statistics literature, this approach is often called recoding), which unlike other existing techniques, such as scrambling or swapping, preserve the truthfulness of the information.

K-anonymity is a technique that prevents joining attacks by generalizing and/or suppressing portions of the released microdata so that no individual can be uniquely distinguished from a group of size k. There are a number of models for producing an anonymous table. One class of models, called global-recoding (Adam et al. 1989) map the values in the domains of quasi-identifier attributes to other values. This chapter is primarily concerned with a specific global-recoding model, called full-domain generalization. Full-domain generalization was proposed by Samarati and Sweeney (Samarati, 2001; Sweeney, 2002) and maps the entire domain of each quasi-identifier attribute in a table to a more general domain in its domain generalization hierarchy. This scheme guarantees all values of a particular attribute in the anonymous table belong to the same domain.

For any anonymity mechanism, it is desirable to define some notions of minimality. Intuitively, a k-anonymous table should not generalize, suppress, or distort the data more than is necessary to achieve such k-anonymity. Indeed, there are a number of ways to define minimality. One notion of minimality is defined so as to generalize or suppress the minimum number of attribute values in order to satisfy a given k-anonymity requirement. Such a problem is shown to be NP-hard (Meyerson et al. 2004). As to our model, the notion of minimal full-domain generalization was defined in (Samarati, 2001; Sweeney, 2002) using the distance vector of the domain generalization. Informally, this definition says that a full-domain generalized private table PT is minimal if PT is k-anonymous, and the height of the resulting generalization is less than or equal to that of any other k-anonymous full-domain generalization.

In this chapter, we focus on this specific global-recoding model of k-anonymity. Our objective is to find the minimal k-anonymous generalization (table) under the definition of minimality defined by Samarati (Samarati, 2001). By introducing the hash-based technique, we provide with a new privacy hash table structure to generate minimal k-anonymous tables that not only improves the search algorithm proposed by Samarati (Samarati, 2001) but is also useful for computing other optimal criteria solution for k-anonymity. Further, we also extend our algorithm to cope with l-diversity.

RELATED WORK

Protecting anonymity when publishing microdata has long been recognized as a problem, and there has been much recent work on computing k-anonymity for this purpose. The u-Argus system (Willenborg, 1996) was implemented to anonymize microdata but considered attribute combinations of only a limited size, so the results were not always guaranteed to be k-anonymous.

In recent years, numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression. The framework was originally defined by Samarati and Sweeney (Samarati, 2001; Sweeney, 2002). Sweeney proposed a greedy heuristic algorithm for full-domain generalization ("Datafly") (Sweeney, 2002). Although the resulting generalization is guaranteed to be k anonymous, there are no minimality guarantees. Samarati proposed the binary search algorithm for discovering a single minimal full-domain generalization. LeFevre et al. described an efficient search algorithm called Incognito, for anonymous full-domain generalization (LeFevre et al. 2005). The concept of l-diversity is introduced by Machanavajjhala et al. in (Machanavajjhala, et al. 2006) to prevent attackers with background knowledge. In (Li, et al. 2007), distribution of sensitive attributes is first considered. Based on this, a more robust privacy measure (which we

15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/privacy-hash-table/61498

Related Content

Policies for Web Security Services
Konstantina Stoupaand Athena Vakali (2006). *Web and Information Security (pp. 52-72).*
www.irma-international.org/chapter/policies-web-security-services/31082

Large-Scale Data Streaming in Fog Computing and Its Applications
Oshin Sharmaand Anusha S. (2021). *Large-Scale Data Streaming, Processing, and Blockchain Security (pp. 50-65).*
www.irma-international.org/chapter/large-scale-data-streaming-in-fog-computing-and-its-applications/259464

Malware Detection by Static Checking and Dynamic Analysis of Executables
Deepti Vidyarthi, S.P. Choudhary, Subrata Rakshitand C.R.S. Kumar (2017). *International Journal of Information Security and Privacy (pp. 29-41).*
www.irma-international.org/article/malware-detection-by-static-checking-and-dynamic-analysis-of-executables/181546

A Survey on Insider Attacks in IAAS-Based Cloud
(2019). *Detection and Mitigation of Insider Attacks in a Cloud Infrastructure: Emerging Research and Opportunities (pp. 28-51).*
www.irma-international.org/chapter/a-survey-on-insider-attacks-in-iaas-based-cloud/221681

Energy, Reliability, and Trust-Based Security Framework for Clustering-Based Routing Model in WSN
Mallanagouda Biradarand Basavaraj Mathapathi (2023). *International Journal of Information Security and Privacy (pp. 1-18).*
www.irma-international.org/article/energy-reliability-and-trust-based-security-framework-for-clustering-based-routing-model-in-wsn/315817