Chapter 14
Graph Representation and Anonymization in Large Survey Rating Data

Xiaoxun Sun
Australian Council for Educational Research, Australia

Min Li
University of Southern Queensland, Australia

ABSTRACT

We study the challenges of protecting privacy of individuals in the large public survey rating data in this chapter. Recent study shows that personal information in supposedly anonymous movie rating records is de-identified. The survey rating data usually contains both ratings of sensitive and non-sensitive issues. The ratings of sensitive issues involve personal privacy. Even though the survey participants do not reveal any of their ratings, their survey records are potentially identifiable by using information from other public sources. None of the existing anonymisation principles can effectively prevent such breaches in large survey rating data sets. We tackle the problem by defining a principle called (k, ε)-anonymity model to protect privacy. Intuitively, the principle requires that, for each transaction t in the given survey rating data T, at least \((k - 1)\) other transactions in T must have ratings similar to t, where the similarity is controlled by \(\epsilon\). The \((k, \epsilon)\)-anonymity model is formulated by its graphical representation and a specific graph-anonymisation problem is studied by adopting graph modification with graph theory. Various cases are analyzed and methods are developed to make the updated graph meet \((k, \epsilon)\) requirements. The methods are applied to two real-life data sets to demonstrate their efficiency and practical utility.

DOI: 10.4018/978-1-61350-053-8.ch014
INTRODUCTION

The problem of privacy-preserving data publishing has received a lot of attention in recent years. Privacy preservation on relational data has been studied extensively. A major type of privacy attack on relational data includes re-identifying individuals by joining a published data set containing sensitive information with the external data sets modeling background knowledge of attackers (Sweeney et al. 2002, Machanavajjhala et al. 2006). Most of the existing work is formulated in contexts of several organizations, such as hospitals, publishing detailed data (also called microdata) about individuals (e.g. medical records) for research or statistical purposes.

Privacy risks of publishing microdata are well-known (Kifer et al. 2006, Wang et al. 2006, Zhang et al. 2008). Famous attacks include de-anonymisation of the Massachusetts hospital discharge database by joining it with a public voter database and privacy breaches caused by AOL search data (Sweeney et al. 2002, Machanavajjhala et al. 2006). Even if identifiers such as names and social security numbers have been removed, the adversary can use linking (Sweeney et al. 2002), homogeneity and background attacks (Machanavajjhala et al. 2006) to re-identify individual data records or sensitive information of individuals. To overcome the re-identification attacks, the mechanism of k-anonymity was proposed (Sweeney et al. 2002). Specifically, a data set is said to be k-anonymous if, on the quasi-identifier (QID) attributes (the maximal set of join attributes to re-identify individual records), each record is identical with at least (k − 1) other records. The larger the value of k, the better the privacy protection is. Although k-anonymity has been well adopted, Machanavajjhala et al. 2006 showed that a k-anonymous data set may still have some subtle but severe privacy problems due to the lack of diversity in sensitive attributes. Particularly, a large body of research contributes to transforming a data set to meet a privacy principle (k-anonymity (Samarati 2001), l-diversity (Machanavajjhala et al. 2006), (α, k)-anonymity (Wong et al. 2006), t -closeness (Li et al. 2007)) using techniques such as generalization, suppression (removal), permutation and swapping of certain data values while minimizing certain cost metrics (Wang et al. 2004, Bayardo et al. 2005, Fung et al. 2005, LeFevre et al. 2005, He et al. 2009).

Recently, a new privacy concern has emerged in privacy preservation research: how to protect the privacy of individuals in published large survey rating data. For example, movie rating data, supposedly to be anonymized, is de-identified by linking un-anonymized data from another source. On October 2, 2006, Netflix, the world’s largest online DVD rental service, announced a $1-million Netflix Prize for improving their movie recommendation service. To aid contestants, Netflix publicly released a data set containing 100,480,507 movie ratings, created by 480,189 Netflix subscribers between December 1999 and December 2005. Narayanan and Shmatikov have shown that an attacker only needs a little bit information of an individual to identify the anonymized movie rating transaction of the individual in the data set. They re-identified Netflix movie ratings using the Internet Movie Database (IMDb) as a source of auxiliary information and successfully identified the Netflix records of known users, uncovering their political preferences and other potentially sensitive information. In this chapter, we will refer to two types of data as “survey rating data” and “relational data”.

MOTIVATION

The structure of large survey rating data is different from relational data, since it does not have fixed personal identifiable attributes. The lack of a clear set of personal identifiable attributes makes the anonymisation challenging (Ghinita et al. 2008, Xu