Chapter 4

A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes[1]

Tamas S. Gal
University of Maryland Baltimore County (UMBC), USA

Zhiyuan Chen
University of Maryland Baltimore County (UMBC), USA

Aryya Gangopadhyay
University of Maryland Baltimore County (UMBC), USA

ABSTRACT

The identity of patients must be protected when patient data is shared. The two most commonly used models to protect identity of patients are L-diversity and K-anonymity. However, existing work mainly considers data sets with a single sensitive attribute, while patient data often contain multiple sensitive attributes (e.g., diagnosis and treatment). This chapter shows that although the K-anonymity model can be trivially extended to multiple sensitive attributes, L-diversity model cannot. The reason is that achieving L-diversity for each individual sensitive attribute does not guarantee L-diversity over all sensitive attributes. The authors propose a new model that extends L-diversity and K-anonymity to multiple sensitive attributes and propose a practical method to implement this model. Experimental results demonstrate the effectiveness of this approach.

INTRODUCTION

Patient data is often shared for research and disease control purposes. For example, the Center for Disease Control and Prevention has a National Program of Cancer Registries which collects data on cancer patients. Such data is made available to public health professionals and researchers to understand and address the cancer burden more effectively.

Privacy is one of the biggest concerns in sharing patient data because without appropriate protection, personal information is vulnerable to misuse. For example, identity theft remains the top concern among customers contacting the Federal Trade Commission (Federal Trade Commission,

DOI: 10.4018/978-1-61692-000-5.ch004

2007). According to a Gartner study (Gartner Inc., 2007), there were 15 million victims of identity theft in 2006. Another study showed that identity theft cost U.S. businesses and customers $56.6 billion in 2005 (MacVittie, 2007). Therefore, legislation such as the Health Insurance Portability and Accountability Act (HIPAA) requires that health care agencies protect the privacy of patient data. This chapter focuses on models that protect identity of patients and at the same time still allow analysis to be conducted on the sanitized data.

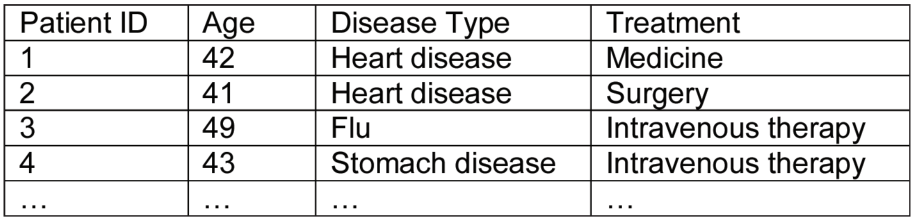
K-Anonymity and L-diversity privacy protection model: The two most commonly used privacy protection models for identity protection are K-anonymity (Sweeney, 2002b) and L-diversity (Machanavajjhala et al., April 2006). K-anonymity prevents *linking attack*,which recovers private information by linking attributes such as race, birth date, gender, and ZIP code with publicly available data sets such as voter's records. Such attributes that appear in both public and private data sets are called *quasi-identifiers*.The K-anonymity model divides records into groups with sizes ≥ K such that each group has identical value or range on quasi-identifier attributes.

Example 1. Figure 1 shows some patient records, where age is the quasi-identifier and disease type and treatment are sensitive attributes (i.e., attributes with privacy sensitive information). Figure 2 shows the anonymized data where the first four rows belong to the same group and have the same range of age. Linking attack cannot discover the identity of a patient using the age attribute because there are at least K (K = 4) patients with the same age range.

L-diversity further enhances K-anonymity by preventing another type of privacy attack called *elimination attack* (which was used by Sherlock Holmes to solve mysteries by excluding the impossible). We use an example to illustrate elimination attack. In Figure 2, if K=3, then the first 3 patients satisfy 3-anonymity. However, they have only 2 different disease type values: heart disease and flu. If someone knows that the patient with ID 3 is unlikely to have heart disease, then he can infer that the patient most likely has flu.

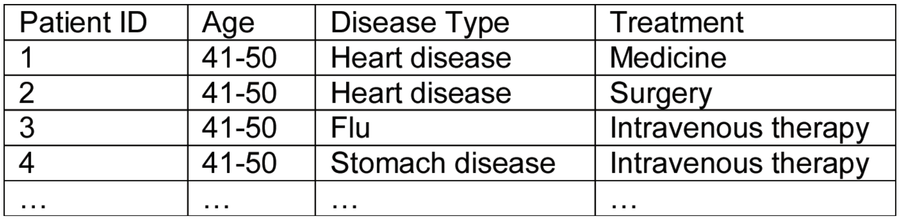
L-diversity prevents elimination attack by requiring that the values of privacy sensitive attributes (e.g., the attribute disease type) in a group have enough degree of diversity. Several

Figure 1. Original patient data



Patient ID	Age	Disease Type	Treatment
1	42	Heart disease	Medicine
2	41	Heart disease	Surgery
3	49	Flu	Intravenous therapy
4	43	Stomach disease	Intravenous therapy
…	…	…	…

Figure 2. Anonymized patient data with K=4



Patient ID	Age	Disease Type	Treatment
1	41-50	Heart disease	Medicine
2	41-50	Heart disease	Surgery
3	41-50	Flu	Intravenous therapy
4	41-50	Stomach disease	Intravenous therapy
…	…	…	…