International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(1): 41-49
The Society of Digital Information and Wireless Communications, 2012 (ISSN: 2305-0012)

Privacy in Medical Data Publishing

Lila Ghemri and Raji Kannah

Department of Computer Science, Texas Southern University
3100 Cleburne Street, Houston, TX 77004
Ghemri_lx@tsu.edu
rajikannah@tsu.edu

ABSTRACT:

Privacy in data publishing concerns itself with the problem of releasing data to enable its study and analysis while protecting the privacy of the people or the subjects whose data is being released. The main motivation behind this work is the need to comply with HIPAA (Health Insurance Portability and Accountability Act) requirements on preserving patient's privacy before making their data public. In this work, we present a policy-aware system that detects HIPAA privacy rule violations in medical records in textual format and takes remedial steps to mask the attributes that cause the violation to make them HIPAA-compliant.

KEYWORDS

Data publishing, privacy, medical, HIPAA.

1. DATA PUBLISHING

Publishing public or semi public data and records offers tremendous benefits to many fields. It allows government agencies to predict and plans for future needs, it allows scientists to develop models of the data being observed, find patterns and connections between attributes and advance their respective fields to realize benefits for the whole humanity. Furthermore with the advent of the Web and the networked world, huge amounts of data are being stored in databases and becoming increasingly available to study, mine and analyze. These advances, however, do not come without a price. Indeed, making these data available, also uncovers vulnerabilities as people private information about themselves, their health, their shopping habits becomes public and the cozy anonymity of being a record amongst thousands suddenly singles out a specific person with their name, age, address all disclosed [1]. This paper is organized as follows: Section 2 presents an overview of the methods used for privacy protection of datasets. Section 3 talks about the laws and regulations in effect to protect medical data. In section 4, we present PACS, a Privacy Aware and System. Section 5 in presents the k-means algorithm and its use in testing the utility of our approach; we will also discuss the results and conclude in section 6.

2. PRIVACY IN DATA PUBLISHING

Privacy in data publishing concerns itself with the problem of releasing data to enable its study and analysis while protecting the privacy of the people or the subjects whose data is being released. Rastogi *et al* [2] give the following definition of privacy in data publishing: "given a database instance containing sensitive information, "anonymize" it to obtain a view such that on one hand attackers cannot learn any sensitive information from the view, and on the other hand legitimate users can use it to compute useful statistics". These two goals may seem to be opposite and contradictory, too little anonymization and the sensitive data can be reconstructed by the attacker,