2006

Database anonymization services

Chad B. Meador
West Virginia University

Database Anonymization Services

Chad B. Meador

Problem Report submitted to the
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Computer Science

Approved By

Bojan Cukic, Ph.D., Chair
John Atkins, Ph.D.
Cindy Tanner

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2006

Keywords: Biometrics, Database Anonymization, Database Security, De-Identification.

Abstract

Database Anonymization Services

Chad B. Meador

The progress of technology and the development of powerful databases have made it possible to store and easily access continually increasing amounts of sensitive data about people. Since personal information is becoming common in many different databases, it is vital that this data be hidden to ensure privacy of the individuals whose records are stored in these repositories. Database anonymization is the key to securing these databases by ensuring that database users will be unable to reveal sensitive personal information by intelligently structuring their queries.

We analyzed the structure of the *BiomData* database which contains images and sound recordings of six biometric modalities acquired from hundreds of volunteers. To ensure the confidentiality of these volunteers, our goal was to prevent queries which would allow database users to obtain images of easily identifiable biometric data (facial images, for example) together with the corresponding images of modalities for which user's anonymity is required (fingerprint images, for example). ERUCES Tricryption® Engine was used to anonymize the links between the six biometric modality tables contained in the database, thereby enhancing privacy of volunteers who participate in the biometric collection study while promoting an open data sharing research environment.

To my family.

Acknowledgements

- *To Dr. Bojan Cukic* – I would like to express my sincere appreciation to my advisor and committee chair for providing me with his time, patience, encouragement, knowledge, and valuable insight. Without your constant guidance my graduate academic career would not have been nearly as fulfilling. I can not thank you enough for being such a great mentor.

- *To Dr. John Atkins and Ms. Cindy Tanner* – I would also like to thank my other committee members for aiding in my graduate studies. I appreciate your time, understanding, and patience in helping me to finalize my graduation. You were both so helpful in working with me during the summer months to hold my defense and make time to meet with me to sign the seemingly never ending stacks of paper work.

- *To Dr. Afzel Noore and Brian Powell* – Thank you so much for allowing me the opportunity to teach CS101. This was a truly enjoyable experience. It is an amazing feeling to know that I have hopefully affected some student's lives in the same way that my favorite instructors have affected mine. Thank you both for making this possible for the past two years of my life.

- *To Nick Bartlow and Robert Goff* - It was great getting to know you guys this year and working together on this project. I wish you the best of luck on your continuing work on this project and with all of your future endeavors.

- *To My Parents* – I would like to offer my vast appreciation and gratitude to my parents for raising me to value my education and always follow through with goals that I set out to achieve. Mom and Dad, you have been my true motivation throughout life. I have accomplished so much at such a young age and know that this would have never been possible without the endless support, guidance, and advice that you have given me. I promise to continue making you proud long after I leave WVU. Thanks for everything.

- *To Laura Vest* – And finally, I would like to thank my girlfriend for always being there to support me, to laugh with me through the fun times, and to help me through the hard times. Without you my life is incomplete and everything that I have accomplished would most certainly not have come as easily as it did. You are always there for me no matter what the situation, and I promise to always do the same for you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

With the progress of commerce, credit, and legal and medical services, the need to maintain information about people is a mandatory part of doing business. All modern governments and businesses must be able to keep detailed records about individuals, which entails storing information about them in some type of database. The progress of technology and the development of very powerful databases have made it possible to store and easily access continually increasing amounts of data about people. This data often includes very sensitive information such as social security numbers, one of the most popular identifiers utilized. Some simpler databases for general commercial uses may store little or none of this sensitive information; however, databases dealing with banking, credit reporting, healthcare, or legal and tax services often contain very large amounts. Since this personal information is becoming so common in many different databases, it is extremely important that this sensitive data be hidden in some way in

order to ensure privacy to the individuals whose information is residing in them. Database anonymization is the key to securing these databases by making certain that an individual who has access to the database will be unable to also gain access to sensitive personal information about the subjects stored within.

1.2 Government Regulations

The government has been taking an active role in attempting to help ensure protection of individual's privacy. It is extremely important that the government creates regulations and laws that help protect this privacy because it is not uncommon for businesses to turn a blind eye to some security issues for the simple sake of not spending the money to remedy the problem, and avoid the loss of business and hindered reputation that occurs as a result of a known security breach to a company's computer systems. Some recent regulations cited by Khulusi are HIPAA, the Gramm-Leach-Bliley Act, SB-1386, and the Data Protection Act, which are helping to mandate the protection of sensitive government, financial, health, and other information [9].

HIPAA stands for Health Insurance Portability and Accountability Act of 1996. This Act requires the Department of Health and Human Services (HHS) to establish national standards for electronic health care transactions and national identifiers for providers, health plans, and employers. It also addresses the security and privacy of health data. Adopting these standards will improve the efficiency and effectiveness of the nation's health care system by encouraging the widespread use of electronic data interchange in health care [14].

The Gramm-Leach-Bliley or GLB Act is also known as the Financial Modernization Act of 1999. The main provisions of this act are to protect the personal financial information of consumers that is held by financial institutions. Financial institutions in this context refers to organizations such as banks, security firms, insurance companies, and basically any companies providing some type of financial product or service to consumers.

SB-1386 is a regulation that became effective in California in July 2003. SB-1386, also known as the Security Breach Information Act, requires the public disclosure of computer security breaches in which private information of any California resident may have been compromised. This law also requires reasonable security controls around California resident data within organizations and their business partners.

The Data Protection Act, which is a regulation in Europe, regulates how personal information is used and protects from misuse of personal details. This act is hinged on eight basic principles that an individual's personal information is: fairly and lawfully processed; processed for limited purposes; adequate, relevant, and not excessive; accurate; not kept longer than necessary; processed in accordance with your rights; kept secure; and is not transferred abroad without adequate protection [8].

1.3 Privacy Threats

Privacy is the ability of an individual or group to stop information about themselves from becoming known to people other than those they choose to give the information to. Privacy is sometimes related to anonymity although it is often most highly valued by people who are publicly known. Privacy can be seen as an aspect of security—one in which trade-offs between the interests of one group and another can become particularly clear [15].

It has become quite clear that numerous threats to privacy can arise from the rather large amounts of personal data that government and commercial organizations are gathering and storing about individuals today. These threats exist because repositories where data reside are vulnerable not only to dangerous hackers with malicious intent, but also to insider abuse. Identity theft, in particular is a threat that has caught on as one of the hottest new forms of crime arising from advances in modern technology. Identity theft is a major problem because of the degree of trouble it can cause to a person. Aside from the

feeling of violation of one's privacy, recovering from identity theft can be a very time-consuming, expensive, and frustrating experience that was caused through no fault of the individual whose identity was stolen.

Data theft is another large threat to privacy. Information can be stolen from a database with the intentions of such things as financial gain from, discrimination against, or gaining advantage over the person or group whose information was stolen, depending on the type and amount of information that is stolen.

Prevention of data tampering and maintenance of data integrity are two other important threats to privacy that need to constantly be addressed. Data tampering can be a threat to a person's financial well-being in the case that a hacker accessed an individual's bank account and illegally transferred funds from their account. This could also affect a person's health in the event that a hacker were to break into a healthcare database and modify such information as prescription types and dosages leading to an innocent individual possibly taking medications that could lead to serious medical problems. Maintaining data integrity is also important as data that is modified or deleted through result of malicious activity or some type of system error or failure can also lead to major problems for an individual who that data is referencing.

The large amounts of data that are being used and shared are going to continue to increase along with the number of databases that are being used to store this data. These databases will always be targets for computer criminals and as such these issues of identity theft, data theft, data tampering, and data integrity are serious threats to privacy, and need to constantly be addressed in the security standards and protocols of database systems.

1.4 Database Security Problems

As previously discussed, databases may store large amounts of sensitive, personal identifying information. The problem occurring is that most databases are vulnerable to attacks which can lead to serious issues such as identity theft. In addition to this vulnerability of most systems to outsider attacks, even the most secure databases that prevent these outsider attacks remain vulnerable to insider misuse and theft [9].

A typical person in today's society does business with several different organizations that store their personal information. This includes all information maintained by the government, financial institutions, lawyers, hospitals, doctors, insurance companies, employers, and even retailers and eCommerce companies. With an individual conducting business with all of these organizations, that person can have sensitive information stored possibly in hundreds of different databases. The threat to this person's sensitive information is as high as the likelihood of the least secure of these database repositories being broken into. In simpler terms this means that it doesn't matter that a person's sensitive information is secure in most databases that it resides in. If that person has information stored in a database that is very prone to a malicious attack then it does them no good that their data is safe everywhere else. The least secure of these databases is the weakest link in the chain and the threat to a person's information is only as strong as this weakest link. Since it is not a feasible task to ensure that all databases that store personal information are at a certain level of security, another solution must be devised to address this problem.

Another problem that exists is that electronic data is constantly released and must meet privacy demands of the individuals whose personal information is being disclosed by the process. Naive approaches to de-identifying these database records can lead to attacks that combine the data with other publicly available information to re-identify the records and discover the identity of the individuals represented. For example, consider a dataset of patient diagnoses that has had all personal identifiers removed. The records in the dataset do not contain a single identifying value; however, many of the records are likely

to contain unique value combinations. Take for instance, a represented individual who is the only male born in 1930 living in a city with a small population. This individual's age, gender, and zip code could be joined with a publicly available voter registry from the area to obtain the person's name, thus revealing their medical history [2].

1.5 The Solution: Database Anonymization

Database anonymization is a solution that can to eliminate threats such as identity theft, data theft, and data tampering while ensuring individual privacy and data integrity. Anonymization is used in databases to de-identify sensitive attributes from their corresponding identifiers. In addition, anonymization helps to ensure privacy to information that has had personal identifiers removed and has been made publicly available, but still faces the possibility of being combined with other publicly available information to re-identify carelessly de-identified records. Anonymization can be done in various ways using various methods depending upon the type of database being anonymized and the actual motivation behind the anonymization.

1.6 *BiomData* Oracle Database

Biometrics is an interdisciplinary field involving personnel from varied organizations ranging from statistics to computer science to law enforcement agencies such as the CIA and FBI. There is currently no centralized system that is available to store data in a secure, reliable, and organized fashion for research use. The *BiomData* Database has been constructed with the goal of providing the necessary biometric dataset for the research community [3]. This is the database that we will be performing anonymization on to de-identify different biometric modalities so that the non-identifiable modalities (i.e. iris, fingerprint, palm print) can not be linked to the identifiable ones (i.e. face,

voice). This will be done through the use of Tricryption which is discussed in detail in Chapter 2 of this document.

1.7 System Features

The *BiomData* database is being specially designed to meet the needs of personnel in the field of biometric research and testing and it possesses many special attributes making it unique including the following:

- Multi-Modal
- Centralized
- Reliable, Flexible, and Accurate
- Web-Enabled
- Time Series Biometric Data
- Template Aging or Biometric Aging
- Statistical Data
- Processed Data and Reusability (Templates, Match Scores)
- High Security

The *BiomData* database is designed to provide data for the community of researchers in the biometric field. This version of the database stores images and sound recordings for six distinct biometric modalities: face, fingerprint, hand geometry, iris, voice and palm. Biometric data can be shared among researchers from different organizations. Data is stored per organization and collection. Collections are divided into sessions and sessions are divided into records. Images for different biometric modalities are stored in different tables. The design allows using multiple scanners for one biometric modality in the same session for the same subject. The design also allows both single captures and also time series captures. In order to keep data in the database consistent, some database fields are restricted to certain values [3].

1.8 Our Contribution

It is extremely vital that a methodology be developed for secure sharing of the data contained in the *BiomData* biometric test database. This methodology needs to include data confidentiality, data integrity, non-repudiation of the data, automated de-identification, and accountability of testers and test organizations accessing the database. Our task in better securing the *BiomData* database focuses mainly on ensuring confidentiality of the biometric data for all subjects stored in the database.

The *BiomData* database includes six separate tables containing the data for six different biometric modalities including face, finger, hand, palm, iris, and voice. The database also contains a record table which acts as a bridge between the six biometric modality tables and the rest of the information related to the subjects in the database. Our task involves anonymizing the links between these tables to ensure confidentiality of the subjects in the database. Confidentiality will be ensured because when the tables are anonymized, users will not be able to recover the true assignment of non-identifiable features, such as iris, fingerprint, and palm print, to identifiable features, such as face and voice. The de-identification of these records will minimize the risks of individualizing images in different modalities, thereby enhancing privacy for data sharing and simplifying security policies and administration.

Chapter 2

Related Work

2.1 ERUCES Tricryption® Engine

To aid us in this project, ERUCES, Inc. allowed us to use the Tricryption® Engine software suite for our experimental setup. The ERUCES Tricryption® Engine is a patent-pending encryption solution that is used to secure sensitive information stored in a database. This is accomplished using standard algorithms along with key management to protect data from theft and tampering, ensuring privacy and integrity in the database. Tricryption claims to eliminate intruder and insider threats to the databases security.

The Tricryption® Engine encrypts individual fields, records, or objects within a database using unique, variable lifetime keys. These keys are stored in a protected database in a separate domain keeping them securely away from the encrypted data. The links between the encrypted data and the corresponding keys are also encrypted.

The full process is detailed below [5]:

- Sensitive data to be encrypted is selected by the user, and a request for encryption is sent
- A randomly generated, transaction-based symmetric key is created; and a random Key ID is created
- The key is encrypted
- The encrypted key and its Key ID are stored in a Key Database
- The Key ID is encrypted, producing a Hidden Link
- The data is encrypted
- The encrypted data and the Hidden Link are returned to the user
- The encrypted data and the key used to encrypt it are completely separated, both physically and logically, and the link between them is hidden.

Tricryption provides many advantages for data-at-rest security. First, when Tricryption is used, the protected data and the encrypted keys are stored separately. Next, Tricryption uses an unlimited number of keys based on the number of transactions as opposed to using a single key or a fixed number of keys. In addition, the complexity of the scrambling system used in Tricryption increases with every single transaction, thus adding to the difficulty of a successful crypto attack. And finally, even if a malicious user were to steal the information database and the key database, the protected data is useless outside of the ERUCES system. The reason for this is the combination of the unlimited number of keys and the complexity of the scrambling system used. Even though the malicious user has the keys used to encrypt the data, there are an unlimited number of them so the user would have to first somehow figure out which keys were used to encrypt which data, then figure out the algorithm used to encrypt the data so he or she could in turn decrypt it.

The following are security features incorporated in the ERUCES Tricryption Engine:

- Standard crypto algorithms including 3DES, AES, and RC4.
- Smart monitoring system with alerts.

- Secure key audit log with reporting.
- Validation of integrity of encrypted data and keys during decryption.
- Accepts requests from only registered and authenticated components.
- Background digital certificates verification.
- Local/Domain Certificate Authority (CA).

2.2 Dedicated Identity Database

Khulusi proposes a method of database anonymization with the underlying principle of transferring the threat to individuals' privacy from the weakest databases that contain their personal information to the strongest and most secure database which can be referred to as a Dedicated Identity Database. The strongest and most secure database can be defined as the database which stores valuable information, personal identifying information in this case, and grants access to authorized individuals, with a legitimate business need, for the use of such data, but prevents anyone else from using, stealing or tampering with such data. In particular, this database must specifically prevent insider misuse of the data. The malicious insider problem is the most difficult problem to solve because insiders, like database administrators, programmers, systems administrators and engineers typically have unlimited access to the raw data. The hacker or malicious outsider problem can also be boiled down to a malicious insider problem. In the event a hacker somehow gains administrator privileges it doesn't matter if it is an administrator or a hacker because after the hacker has breached the database and attained administrator privileges, he or she is basically an insider while accessing the database [9]. The weakest databases can in turn be defined as the databases which use less security features and are more prone to attacks and insider abuse.

Transferring the threat to an individual's privacy from the weakest databases that contain their personal information to the strongest and most secure database is done by only storing personal identifying information in the Dedicated Identity Database meaning that the threat of an attacker or insider accessing and possibly tampering with or stealing a

person's personal information will reduce significantly because it resides only in the most secure database. In this anonymization of databases, the Dedicate Identity Database will contain all sensitive personal information and other databases that may or may not be less secure will hold any other information that is not sensitive. In this manner a hacker or malicious insider may be able to access in a database such information as transaction data, medical records, items purchased or sold, or tax information to name a few, but will not be able to see the personal data related to it thus making the data they can see useless. An example of this given by Bassam Khulusi is "what good is it to know that somebody was treated for a psychological illness, when there is no name, social security number, address, or anything of a personal nature attached to that information [9]." With this lack of access to the personal information, identity theft is not possible. Data tampering would be pointless unless it were being performed to simply "vandalize" a database, but still could not be targeted at a specific known individual in the database.

In Khulusi's Dedicated Identity Database all personal identifying information is stored and is accessed by the weaker and less secure databases that have access to the data only on a need to know basis. A pseudo-identity is created in the less secure databases which is a cryptographically generated pseudonym used for accessing each individual identity in the Dedicated Identity Database. This cryptographically generated pseudonym is the only thing relating an individual's sensitive information in the Dedicated Identity Database to any other non-sensitive data in one of the weaker databases. This is the concept of de-identification in which the person's sensitive information is continuously de-identified from their non-sensitive information meaning there is no identity attached to the non-sensitive information. This data in the weaker databases is therefore said to be anonymous because instead of being stored with a person's name, social security number, and other sensitive information, it is stored only with a cryptographically generated pseudonym. If one of the weaker databases does actually need sensitive data for an individual, a request must be sent to the Dedicated Identity Database and only through the submission of this legitimate request could the needed identifying information be obtained [9].

2.3 *K*-Anonymization

K-Anonymization is a technique that is also used to de-identify sensitive attributes from their corresponding identifiers. This anonymization approach is used to address the problems created by the large amounts of information that are publicly available that cause questions to be raised about the protection of individual's privacy. Zhong points out that this particularly includes large datasets containing sensitive information (e.g., healthcare information) that are available for public access and simply have the sensitive identifiers stripped off in an attempt to protect privacy of the individual's whose information is published [16]. A privacy issue is encountered when it becomes possible to use this information, in conjunction with some other type of publicly available database, to associate a person's corresponding identifiers. This was previously discussed in section 1.4 of this document in the example given regarding a public medical database and voter lists being used to find out private information about an individual's medical conditions and history.

A *k*-anonymized dataset has the property that each record is indistinguishable from at least *k*-1 others. Zhong gives the following example to illustrate a *k*-anonymized table where k is equal to 2. For this example, each record will be indistinguishable from at least one other record in the table since the table is 2-anonymized.

Consider a table containing health information of medical studies as shown in Table 1. Each record includes a patient's date of birth, zip code, allergy, and history of illness. Although the identifier of each patient does not explicitly appear in this table, a dedicated adversary may be able to derive the identifiers of some patients using the combinations of date of birth and zip code. For example, he may be able to find that his roommate is the patient of the first row, who has allergy to penicillin and a history of pharyngitis.

Date of Birth	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Table 2.1: Table of Health Data

In this example the set of attributes {date of birth, zip code} is referred to as a quasi-identifier, because these attributes can be used in combination to identify an individual. Attributes such as allergy and history of illness are referred to as sensitive attributes. The privacy threat that is being considered is that the adversary may be able to link the sensitive attributes of some rows to their corresponding identifiers using the quasi-identifiers. The strategy devised to alleviate this threat is to make the table *k*-anonymous.

Date of Birth	Zip Code	Allergy	History of Illness
*	07030	Penicillin	Pharyngitis
08-02-57	0702*	No Allergy	Stroke
*	07030	No Allergy	Polio
08-02-57	0702*	Sulfur	Diphtheria
*	07030	No Allergy	Colitis

Table 2.2: 2-Anonymized Table of Health Data

In this k-anonymous table, each value of the quasi-identifier appears at least *k* times. Zhong states in this example that if the adversary only uses the quasi-identifiers to link sensitive attributes to the identifiers, then each involved entity (patients in this example) is "hidden" in at least *k* peers. Two methods that Zhong points out of achieving this k-anonymization are generalization and suppression which are both used in this example. In the above example suppression is illustrated in the "date of birth" field where some entries are replaced with a "*". Generalization is illustrated in the "zip code" field where some of the entries are only partially displayed (e.g. replacing some or all occurrences of "07028" and "07029" with "0702*").

Chapter 3

Case Study

3.1 *BiomData* Database Structure

As previously stated, the *BiomData* database has been constructed as a centralized system available to the biometric research community, which contains biometric data stored in a secure, reliable, and organized fashion. This is a quite complex database including 27 tables with 34 different relationships linking the tables. There are six separate tables containing the data for six different biometric modalities including Face, Finger, Hand, Palm, Iris, and Voice. There exists a Record table which acts as a bridge between the six biometric modality tables and the rest of the information related to the subjects in the database. And finally, there is a Subject table that contains various sensitive and non-sensitive information about the subject's whose biometric data is stored. This report describes the process of anonymizing the links between the Record table, the Subject table, and the six biometric modality tables to ensure confidentiality of the subjects in the database. Confidentiality needs to be ensured because in the current schema, users are

able to recover the true assignment of non-identifiable features, such as iris, fingerprint, and palm print, to identifiable features, such as face and voice. The de-identification of these records will minimize the risks of individualizing images in different modalities thereby enhancing privacy for data sharing and simplifying security policies and administration.

3.2 Original Database Schema

The original structure of the tables that will be affected by the anonymization of the database is shown in Figure 3.1:

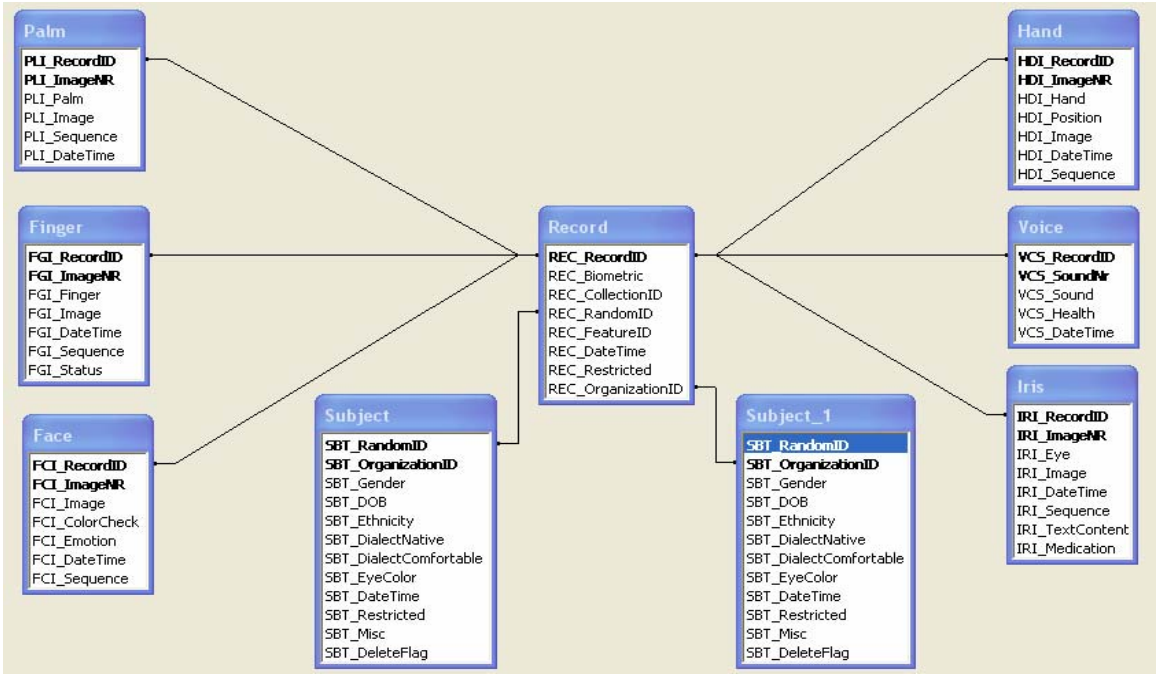

Figure 3.1: Partial View of the Original *BiomData* Database Schema

In Figure 3.1, the bolded fields represent the primary keys in the tables. It can be observed that each modality table contains a pair of fields for the primary key. The first of these fields is a Record ID field which is used to link each record in the modality table to its corresponding record in the Record table. The records in the Record table are then

related to the Subject table through two separate fields, which are used as a composite primary key in the Subject table. These two fields are unique identifiers for each subject in the database. It can also be observed that there are two copies of the Subject table shown. There are not actually two separate Subject tables in the database. This simply illustrates that two separate relationships exist between the Record and Subject tables.

3.3 Table Modifications

In the anonymization process we will first de-identify the relationship between the Record and Subject tables. Next, the Record table will be split into two tables: Record_Spiis and Record_Nspiis. Record_Spiis contains the sensitive data, and Record_Nspiis contains the non-sensitive data (Spii in the table name means sensitive personal identification information and Nspii in the table name means non-sensitive personal identification information). There should be no direct link between the Record_Spiis and Record_Nspiis tables. Figure 3.2 shows the original Record table and the two new tables created after the table is split.

Original Record Table

REC RecordID	REC Biometric	REC CollectionID	REC RandomID	REC FeatureID	REC DateTime	REC Restricted	REC OrganizationID

New Record_Spiis Table

REC RecordID	REC RandomID	REC Restricted	REC OrganizationID

New Record_Nspiis Table

REC EdiHLdi	**REC EriHLdi**	**REC HLri**	REC Biometric	REC CollectionID	REC FeatureID	REC DateTime

Figure 3.2: Record Table Modifications

From the above figure it can be seen that three new fields (shown in bold) are introduced into the Record_Nspiis table. These fields are the hidden links that are used to traverse amongst the split record tables and the six modality tables. These hidden links are

produced by encrypting the randomly generated symmetric keys for each record and encrypting the random key id that is assigned to each key. The fields REC_EdiHLdi and REC_EriHLdi are two different encrypted keys and the HLri is the encrypted key id associated with the REC_EriHLdi key.

Each of the modality tables contains a new field that acts as a hidden link to associate the modality table to the Record_Nspiis table. This field is the encrypted key id associated with the REC_EdiHLdi in the Record Table. These encrypted key id's are completely separated, both physically and logically, the links between them are hidden. This can be observed in Figure 3.3, which illustrates the modifications to the Face table.

Original Face Table						
FCI RecordID	FCI ImageNr	FCI Image	FCI ColorCheck	FCI Emotion	FCI DateTime	FCI Sequence

Modified Face Table						
FCI EdiHLdi	FCI ImageNr	FCI Image	FCI ColorCheck	FCI Emotion	FCI DateTime	FCI Sequence

Figure 3.3: Modality Table Modifications

Each of the modality tables will be modified in the same way as the Face table which is shown above. The field in each modality table that contains the Record ID will be replaced with a new field that contains the hidden link associating the given table to the Record_Nspiis table. These links will make it impossible to associate an identifiable feature to a non-identifiable feature such as a fingerprint to a face scan.

3.4 Modified Database Schema

Upon modification of the database, the six modality tables will contain a hidden link to the associated record in the Record_Nspiis table. The Record_Nspiis table and

Record_Spiis tables will be completely de-identified and hidden links will be contained in both to associate the related records. The Subject table will then only be related to the Record_Spiis table via 2 separate relationships using the two fields that represent the composite key in the Subject table. The new structure of the database can be seen in Figure 3.4, and Tables 3.1-3.3 illustrate a generic view of how the data would look in the tables under the new schema.

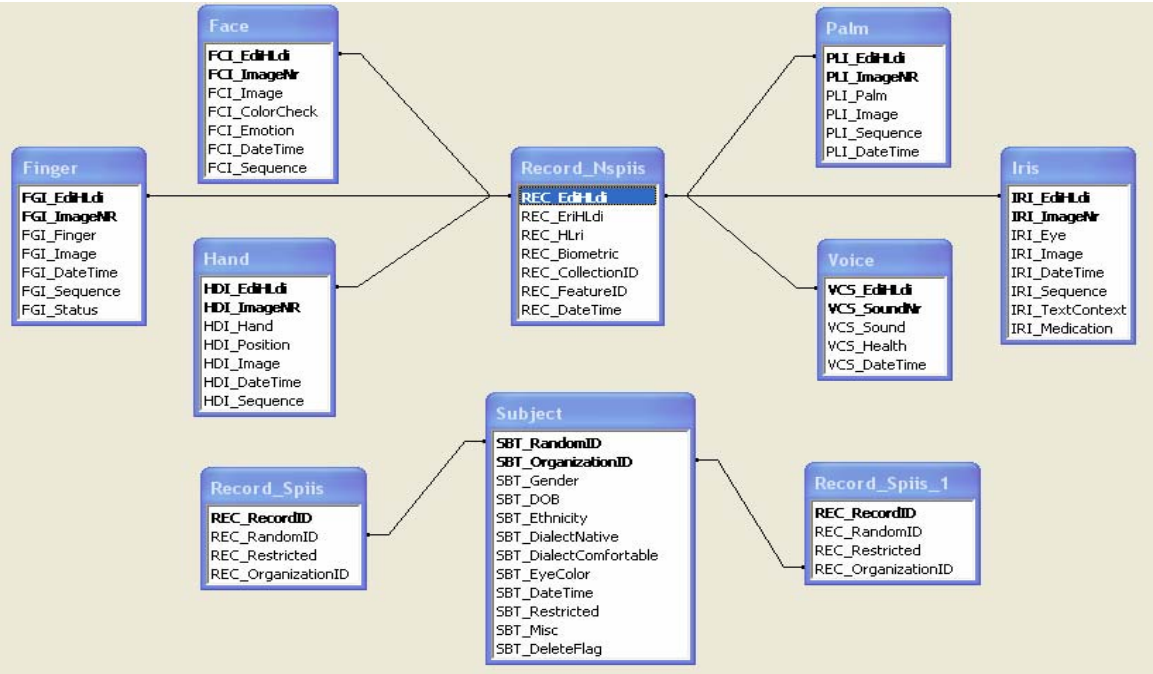

Figure 3.4: Modified Database Schema

RECORD_SPIIS TABLE				
REC RecordID	**REC RandomID**	**REC Restricted**	**REC OrganizationID**	**HL_{Di}**
1	1135462	Y	1	ACEx0aZG1CpDv…
2	2987698	N	2	ACEx0SxzJVANh…
3	1145539	N	3	ACEx0UYDsG91X…
4	2193334	Y	4	ACEx0UgoBXaC7…
5	3490046	N	5	ACEx0eeeer4LB7…
6	2287654	Y	6	ACEx0V2Q0djom…

Table 3.1: New Record_Spiis Table

RECORD_NSPIIS TABLE					
E_{Di}	**E_{Ri}**	**HL_{Ri}**	**REC Biometric**	**REC CollectionID**	**…..**
kXE3sFcE+Jx…	GxINhfUZ/ufCP…	ACEx0aZG1Co…	FC	1	
wK1GT9uWHn…	Cg9YsQ2GRw…	ACEx0SxzJVD…	FC	2	
lDbiOW1L6rZR…	FHqxDUjXRYR…	ACEx0UYDsG9…	FC	3	
ANqN1rmvlHsF…	T7IUNbxY7ci0d…	ACEx0UgoBXY…	FC	4	
EvsI16Z+8fF8l…	PxTFVzalYnt/Y…	ACEx0eeeer52…	FC	5	
KksQHv2yOZP…	8qzR0MtbLuek…	ACEx0V2Q0dg…	FC	6	

Table 3.2: New Record_Nspiis Table

FACE TABLE					
E_{Di}	**FCI ImageNR**	**FCI ColorCheck**	**FCI Emotion**	**FCI Sequence**	**…..**
kXE3sFcE+Jx…	1	Y	Hate	0	
wK1GT9uWHn…	1	Y	Disgust	0	
lDbiOW1L6rZR…	1	N	Surprise	0	
ANqN1rmvlHsF…	1	Y	Unknown	0	
EvsI16Z+8fF8l…	1	N	Other	0	
KksQHv2yOZP…	1	Y	Laughter	0	

Table 3.3: Modified Face Table

Chapter 4

Summary

4.1 Summary

In this day and age, computers are an essential part of everyday life. Subsequently, increasing numbers of computer applications use and store sensitive personal information in numerous databases. Hackers and malicious insiders pose a large threat to individual's privacy because they can possibly gain access to this personal information and use it for very serious crimes such as identity theft. To minimize this threat to privacy, databases need to use anonymization to de-identify an individual's sensitive attributes from any corresponding identifiers. When anonymization is used, a hacker or even an insider could gain access to tables in a database but still not be able to reveal any extra information that can be used to link sensitive attributes to corresponding identifiers. This increase in security is vital to protect the privacy of all individuals who have sensitive personal information stored in any type of database.

In the context of our work, we are currently anonymizing a biometric database containing data for six different biometric modalities including hand geometry, face, voice recognition, iris, palm, and fingerprint. The motivation behind this anonymization is to ensure confidentiality of the individual's whose information is stored in the *BiomData* database. This database will be anonymized through the use of Tricryption to ensure confidentiality of the biometric data for all subjects stored in the database. The anonymization of this database will minimize the risks of individualizing images in different modalities, thereby enhancing privacy for data sharing and simplifying security policies and administration.

Bibliography

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Anonymizing Tables," *10th International Conference on Database Theory*, Jan 2005.

[2] Roberto Bayardo and Rakesh Agrawal, "Data Privacy Through Optimal k-Anonymization," *21st International Conference on Data Engineering*, pp. 217-228, Apr 2005.

[3] Simona Crihalmeanu, "Database Design for "*BiomData*" Oracle Database – Technical Documentation," WVU, Lane Department of CSEE, Jan 2006.

[4] M. Douglass, G. Clifford, A. Reisner, G. Moody, and R. Mark, "Computer-Assisted De-Identification of Free Text in the MIMIC Database," *Computers in Cardiology*, pp. 341-344, Sep 2004.

[5] A. Kalam, Y. Deswarte, G. Trouessin, and E. Cordonnier, "Personal Data Anonymization for Security and Privacy in Collaborative Environments," *International Symposium on Collaborative Technologies and Systems*, pp. 56-61, May 2005.

[6] Andrew Thibault, "Protecting Information Storage: Is Your Data Secure?" Jan 2001, http://www.itaa.org/events/event.cfm?EventID=295.

[7] Bugra Gedik and Ling Lui, "Location Privacy in Mobile Systems: A Personalized Model," 25th *IEEE International Conference on Distributed Computing Systems*, pp. 620-629, June 2005.

[8] Information Commissioners Office, "Data Protection," http://www.ico.gov.uk/eventual.aspx?id=34.

[9] Khulusi, Bassam, "Anonymization Services: A Brief Overview," ERUCES, Inc., http://www.tricryption.com.

[10] Kristen LeFevre, David DeWitt, and Raghu Ramakrishnan, "Mondrian Multidimensional K-Anonymity," *22nd International Conference on Data Engineering*, pp. 25-35. Apr 2006.

[11] M. Ercan Nergiz, and Chris Clifton, "Thoughts on k-Anonymization," *22nd International Conference on Data Engineering Workshops*, pp. 96-105, Apr 2006.

[12] Elaine Newton, Latanya Sweeney, and Bradley Malin, "Preserving Privacy by De-Identifying Face Images," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, issue 2, pp. 232-243, Feb 2005.

[13] Adam Slagell and William Yurcik, "Sharing Computer Network Logs for Security and Privacy: A Motivation for New Methodologies of Anonymization," *Workshop of the 1st International Conference on Network Forensics*, pp. 82-91, Sep 2005.

[14] U.S. Department of Health and Human Resources, "HIPAA – General Information," Dec 2005, http://www.cms.hhs.gov/HIPAAGenInfo/.

[15] Wikipedia, "Privacy," http://en.wikipedia.org/wiki/Privacy.

[16] Sheng Zhong, Zhiqiang Yang, and Rebecca Wright, "Privacy-Enhancing k-Anonymization of Customer Data," *24th ACM Symposium on Principles of Database Systems*, pp. 139-147, 2005.