International Journal of Engineering & Technology, 3 (3) (2014) 337-342
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v3i3.2139
Research Paper

Service oriented architecture and privacy preserving mashup of healthcare data

R. Vijayalakshmi [1*], N. Duraipandian [2]

[1] *Asst. Professor, Dept. of IT Velammal Engineering College*
[2] *Principal Velammal Engineering College*
**Corresponding author E-mail: viswa31999@gmail.com*

Abstract

Mashup of health care data from different medical sources must be privacy preserved since the data recipient and/or the data provider may not always be a trusted party. Raw medical data contains person specific sensitive information like ailment, surgery etc. and hence it is susceptible to certain privacy attacks such as attribute linkage and record linkage. There are different privacy models to thwart the privacy attacks. This paper illustrates how to vertically integrate the data from mental health clinic and National AIDS Control Organization (NACO) and preserve privacy using the LKC privacy model.

Keywords: *Mashup, Linkage, Anonymize Utility, Diversity, Adversary, Interoperability.*

1. Introduction

Mashup service [1] is a web technology to integrate information from two or more data providers. It was introduced in the year 2010. Mashup is also used to integrate data from tables. A medical research is very important as it aims to advance the knowledge on a medical condition and to improve human health. For the purpose of medical research, fields from tables of different centres such as hospitals, pharmacies, insurance companies, government agencies etc. need to be integrated. Currently AIDS related medical research focuses on studies such as:

- What percentage of HIV infected patients in India is diagnosed with suicidal ideation?
- How many HIV infected patients diagnosed with suicidal ideation have died only because of their mental health disorder?

The medical research can also be useful to the Indian government to address the health needs of the patients as there are also patients below the poverty line and qualified doctors and advanced medical therapies are concentrated only at the cities.

1.1. Need for privacy

Every person on earth needs privacy. Each individual has certain information about himself/herself which he/she do not want to disclose to anyone or only to a certain number of people. This information is known as sensitive information. Privacy is needed everywhere, at school, at workplace and in personal life. For example, password of an e-mail account is a sensitive data since it can be used to read the personal/official mails of a person. Everybody has a right to privacy and it is a crime to intrude into the privacy of a person.

In relational databases, a table can contain person-specific sensitive information such as salary, ailment, account balance etc. Trust to one party may not necessarily be transitive to a third party [1]. For example, patients of a hospital trust the hospital and submit their personal data to the hospital. The hospital has to publish the patients data to an external medical centre for statistical or classification analysis. But the trust is not transitive i.e. the patients do not trust the medical centre. Therefore the hospital must ensure that privacy is preserved on the data before releasing the data to the medical centre i.e. the data recipient which is the medical centre in this case must not be able to find out which sensitive information belongs to which individual. In this example, the patients are the record owners since each record is associated with each patient. And the hospital is the data publisher.

Research has shown that the published data is not privacy-preserved if explicit identifying information, such as social security number, name, telephone number and address is removed. Quasi Identifiers (QID) are a set of attributes, which when exploited can reveal the sensitive value of the record owner. Gender, date of birth, and postal code are examples of quasi-identifiers.
The real life example by P. Samarati [2] illustrates the need for privacy. The Group Insurance Company (GIC) of Massachusetts published the medical data in Table 1. To prevent linking attacks, SSN and Name fields were left empty. And later it was found that suppressing these fields alone is not sufficient to ensure privacy. This was discovered when an external table, the Voter list table (Table 2) was published by the state government of Massachusetts. The two tables had common attributes (DateOfBirth, Sex and ZIP). By linking these quasi identifiers to the two tables, individual

Table 1: Medical Data

SSN	Name	Race	Date of Birth	Sex	Zip	Marital Status	Health Problem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

Table 2: Voter List

Name	Address	City	Zip	DOB	Sex	Party
Sue J.Carlson	900 Market St.	SanFrancisco	94142	9/15/61	female	democrat

Privacy was exploited and as a result the medical problem of the governor of Massachusetts was found

1.2. Privacy threats in data mashup

A data mashup application which integrates tables vertically can be misused by adversaries to reveal person-specific sensitive information that was not available before the mashup. For a single table, the privacy threat is only from the data recipient. For a mashup table, the privacy threat can also come from the data provider. For example, let's say the data from two data providers A and B is vertically integrated. Sensitive attribute is only at A. In this case, B can be an adversary trying to discover the value of the sensitive attribute of an individual after the mashup.
The adversary's knowledge of the values in QID attributes can help the adversary in discovering the sensitive information of the individual from the published table. The following types of threats exist today:

1.2.1. Record linkage

Here, the adversary can identify the record of the individual and thereby his sensitive value says Benjamin C.M. Fung [1]. Let's say that the adversary knows that the Job of the person is Engineer and the age of the person is 35. In table 3 there is only one record, the fifth record which has 'Engineer' in job and '35' in Age. So the adversary will identify the record of the individual if he has access to the table and thereby come to know the individual's sensitive value.

1.2.2. Attribute linkage

Unlike record linkage, the exact record of the individual cannot be identified here. But the sensitive value of the individual can be inferred. For example, let's say the adversary knows that the age of the person is 35. In Table 3, the fifth and the seventh records have 35 in 'Age'. So the adversary will not be able to find whether the fifth record or the seventh record belongs to the individual, but he will be able to find the sensitive value of the person since both the records have the same sensitive value.

2. Preventing privacy threats

Anonymization is an approach to thwart privacy threats. It refers to the generalization, suppression etc. of quasi identifiers that seeks to hide the identity and/or the sensitive data of record owners. Privacy models can be used to anonymize the tables

Table 3: Raw Data

Shared	Data Provider A		Data Provider B	
UID	Sensitive	Gender	Job	Age
1	S2	M	Doctor	34
2	S2	F	Engineer	46
3	S2	M	Clerk	54
4	S2	M	Teacher	60
5	S1	F	Engineer	35
6	S2	M	Manager	43
7	S1	F	CEO	35
8	S2	M	Doctor	42
9	S2	F	Lawyer	46
10	S2	F	Salesman	28
11	S2	M	Clerk	32

2.1. Anonymization of a single table

Let's say that a data publisher has a table of the form T(Explicit Identifier, QID, Sensitive Attributes, Non-Sensitive Attributes) where Explicit Identifier is a set of attributes, such as Social Security Number (SSN)and name that can be used to identify record owners explicitly; Quasi Identifier (QID) is a set of attributes such as age, gender , place etc that can be matched to identify an individual's record and/or his/her sensitive value; Sensitive Attributes consists of sensitive person-specific information such as health problem, salary, account balance etc; and Non-Sensitive Attributes are the other attributes that do not fall into the previous three categories. To prevent linking attacks, the data publisher modifies QID in the original table T to QID_ by applying anonymization operations such as generalization, suppression etc. He does not publish the original table. Instead he publishes an anonymous table of the form, T^1 (QID_, Sensitive Attributes, Non-Sensitive Attributes).Privacy models such as k-anonymity, l-diversity, confidence bounding model etc can be used to anonymize the original table. The anonymized table must also retain as much data as possible for information utility. Metrics can be used to measure the utility of an anonymous table. The Non-Sensitive Attributes are not published in all cases. They are published only if they are really needed for analysis.

2.2. Centralized and distributed anonymization in data mashup

Let's take medical data as an example. The Hong Kong Blood Transfusion Service (BTS) collects blood from donors and distributes the blood to the public hospitals in Hong Kong. Each hospital stores details about patients such as SSN, name, age, gender, zipcode, type of surgery done, whether blood was transfused for the surgery etc. The hospitals are required to provide periodically, the blood usage data together with patient specific surgery data to the BTS for data analysis. There are two different ways to provide privacy in data integration. In the centralized approach, a central government health agency can integrate the raw data from different hospitals and then preserve privacy on the integrated data; this approach is also called integrate-then-generalize approach. This approach can be used only if the mashup application is from a trusted party. The distributed approach is used if there is no trusted party. Here, each hospital should not learn more information about other hospitals other than what is present in the final mashup table. This is different from the generalize-then-integrate approach, where-in each data provider anonymizes its table independently before the mashup and so suffers from a good degree of information utility loss.

3. Integration of health care data

A new privacy problem with dynamically integrating the health care data of HIV/AIDS patients from NACO (National AIDS Control Organization) and psychiatric clinics in India is identified [5]. NACO [6] comes under the ministry of health and family welfare. It collects details about HIV/AIDS patients from STD clinics throughout the country. It develops programs for prevention and control of HIV/AIDS in India. A medical research has shown that many People Living with HIV/AIDS (PLHA) also suffer from psychiatric disorders [9], [10] such as anxiety, depression, suicidal ideation etc. It may be due to many reasons such as poor social adjustment, poor family relations, physical pain, the thought of living with the disease etc. Owing to the mental health problems, these patients are less responsive to treatments. The integration of data for joint data analysis will be helpful to NACO to also address the mental health needs of PLHA effectively through social and behavioural strategies such as counselling, social support and psychotherapeutic strategies.

3.1. Integration of the tables from NACO and mental health clinic

Here, NACO is considered to be a trusted party and so we can take the centralized approach in data mashup. NACO can have access to the raw records of HIV/AIDS patients in psychiatric clinic. But psychiatric clinic must not be allowed to access the NACO records of PLHA. NACO will integrate the two tables vertically using AADHAR and will then generalize the integrated table using the LKC privacy model. In the integrated table, AADHAR will be replaced by serial number as shown in Table 4. Here, psychotic disorder and suicidal ideation are the sensitive values. Age, sex and place are the QID attributes. The LKC privacy model is applied to this integrated table to thwart both attribute and record linkages. Table 5 is an example of an anonymous version of Table 4.

3.2. Service oriented architecture

The two data providers, NACO and mental health clinic can be operating on heterogeneous platform and therefore Service Oriented Architecture (SOA) [11] is used to tackle this heterogeneity. SOAP request and response messages will be used to send mental health records of PLHA from the psychiatric clinic to NACO. Since SOAP is based on XML, which is platform independent, we can achieve interoperability between the two data providers as shown in Fig 1. NACO admin can access the integrated raw data but the medical researcher can see only the privacy preserved data.

3.3. LKC privacy model

LKC [1] is a new privacy model developed by Benjamin C.M. Fung. L, K and C are thresholds set by the data provider; in this case it is set by NACO. Older models such as k-anonymity, l-diversity, (α,k) anonymity, confidence bounding model etc. suffer from "curse of high-dimensionality" when used for high dimensional mashup i.e. as the number of attributes increases , more and more generalization is done leading to loss of information utility. LKC model overcomes this curse, by taking advantage of one of the limitations of the adversary - the adversary will not know the values of all the QID attributes. If an adversary knows atmost L values, then atleast K number of records must contain same values in L number of QID attributes. The confidence in inferring a sensitive value from the QID

Table 4: Example Data

1	31	M	Mahabalipuram	psychotic disorder
2	48	M	Marakkanam	alcohol dependence
3	31	M	Chennai	suicidal ideation
4	24	M	Athipattu	depression
5	48	M	Chennai	alcohol dependence
6	46	M	Mahabalipuram	depression
7	4	M	Marakkanam	behavioural disorder
8	48	F	Athipattu	anxiety disorder
9	46	F	Marakkanam	adjustment disorder
10	65	F	Kanchipuram	suicidal ideation
11	65	F	Tindivanam	anxiety disorder

Table 5: Anonymous Data (L = 3, K = 2, C = 0.5)

1	1-60	M	Mahabalipuram	psychotic disorder
2	1-60	M	Coastal_rural_TN	alcohol dependence
3	1-60	M	Chennai	suicidal ideation
4	1-60	M	Coastal_rural_TN	depression
5	1-60	M	Chennai	alcohol dependence
6	1-60	M	Mahabalipuram	depression
7	1-60	M	Coastal_rural_TN	behavioural disorder
8	1-60	F	Coastal_rural_TN	anxiety disorder
9	1-60	F	Coastal_rural_TN	adjustment disorder
10	61-99	F	Noncoastal_urban_TN	suicidal ideation
11	61-99	F	Noncoastal_urban_TN	anxiety disorder

Attributes must not be greater than C and C must always be less than 100% or 1. With the LKC model, the probability of a successful record linkage is limited to $\leq 1/K$ and the probability of a successful attribute linkage is limited to $\leq C$. And LKC is the first model to be used for high dimensional mashup.

To generalize the QID values, a taxonomy tree is built for each of the QID attributes as shown in Fig 2. The top of the tree contains the most generalized value for an attribute and the leaves of the tree contain the most specialized values. As of now, as only one psychiatric hospital is concentrated, a mental health hospital in Chennai is focused upon where people from nearby places visit here and so the most generalized value for place would be Tamil Nadu.

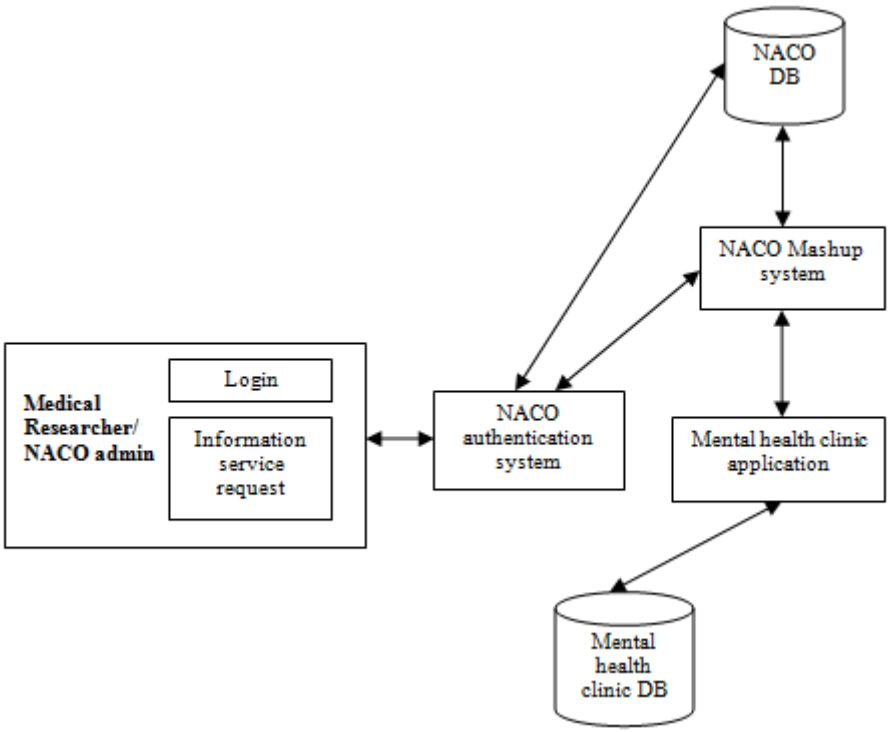

Fig. 1: Architecture for Privacy Preserving Data Mashup

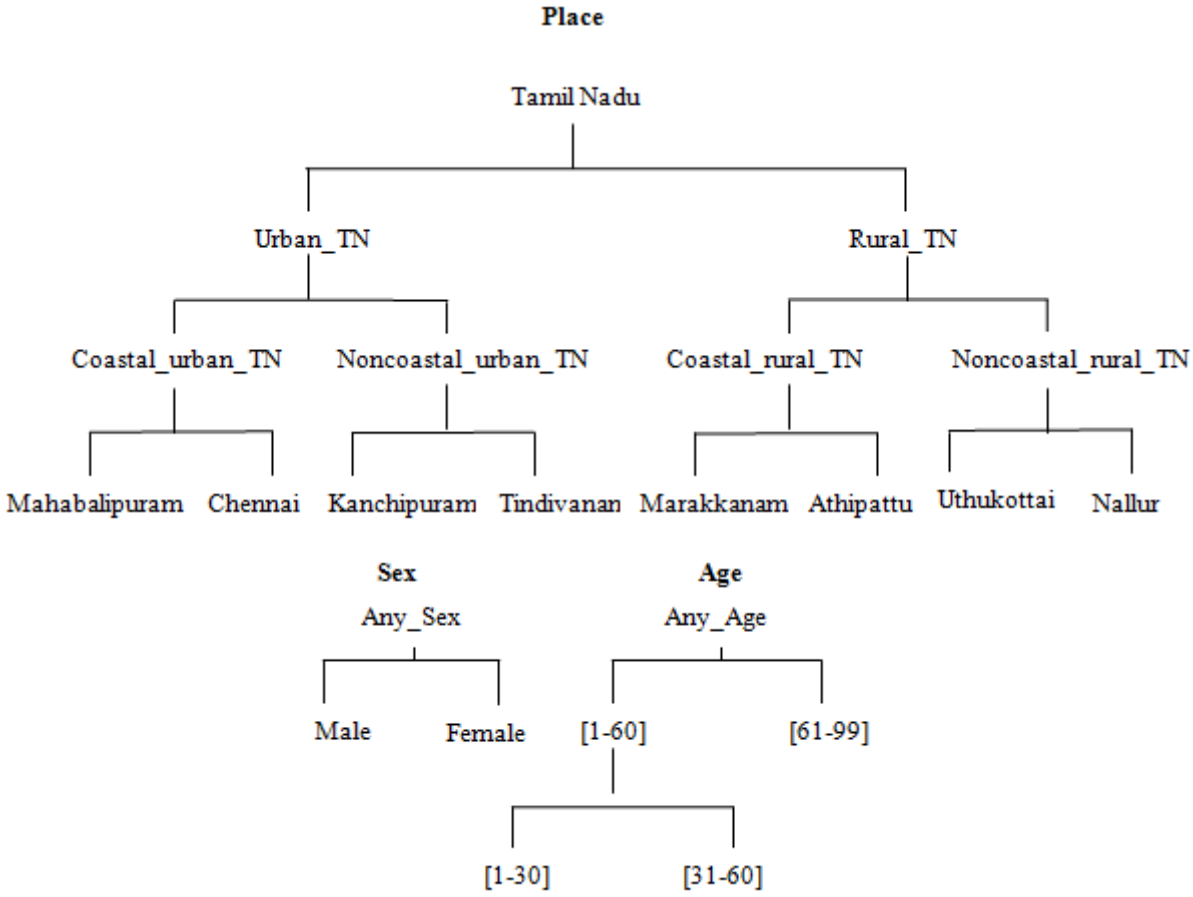

Fig. 2: QID attributes and taxonomy tree

3.4. Utility measure

To achieve a privacy requirement without compensating the information utility, (1) is used to measure the data quality in the anonymous table i.e. (1) is used to measure the amount of data distortion in the anonymous data. And the qid attribute with the maximum Score value is chosen for specialization.

Score (v) = $\sum | T [qid_v] |^2$ (1)

For example, Score (Any_Sex) = $7^2 + 4^2 = 65$ since there are 7 males and 4 females in Table 4.

3.5. Privacy preserving algorithm

A top down specialization algorithm is used to preserve privacy, wherein the initial values of every QID attribute is set to the topmost value in the taxonomy tree and in each iteration, a QID value with the highest score is chosen for specialization. The algorithm is as follows:

1) Initialize every QID attribute (Age, Sex and Place) in table T to the topmost value in the taxonomy tree.
2) Initialize $UCut_i$ to include the topmost value of all the QID attributes
3) I.e. $UCut_i$ = {Any_Sex, Any_Age, Tamil Nadu} Compute Valid(x) for every x € $UCut_{i\ Valid}(x)$ = false if any further specialization of x would lead to a violation of the LKC privacy requirement otherwise true.
4) Compute Score(x) for every x € $UCut_i$
5) while some x € $UCut_i$ is valid do
6) Specialize x that has the maximum Score value
7) Compute Validity for every child(x)
8) Replace x in $UCut_i$ with child(x)
9) Compute Score for the new values in $UCut_i$
10) End while
11) Output $UCut_i$

References

[1] Benjamin C.M. Fung, Thomas Trojer, Patrick C.K. Hung, Li Xiong, Khalil Al-Hussaeni, and RachidaDssouli, "Service-Oriented Architecture for High-Dimensional Private Data Mashup" IEEE Transactions on Services Computing, vol. 5, no. 3, pp. 373-386, July-September 2012.

[2] Latanya Sweeney, "k-Anonymity: A model for protecting privacy" International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, pp.557-570, May 2002.

[3] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, pp. 14:1-14:53, June 2010.

[4] Shankar Das,"High prevalence of mental health problems in HIV/AIDS patients," http://indianexpress.com/news/high-prevalence-of-mental-problems-in-hivaidspatients/896442/, January 2012.

[5] "HIV/AIDS cases," http://mapsofindia.com/census/ hiv-aids- cases-in-india.html.

[6] "National AIDS Control Organization Department of AIDS Control," http://nacoonline.org

[7] "Nabil Adam, Tom White, Basit Shafiq, Jaideep Vaidya, Xiaoyun He, "Privacy Preserving Integration of Health Care Data," http:/ncbi.nlm.nih.gov/pmc/articles/PMC2655922/, AIMA Annual Symposium Proceedings, 2007

[8] M.Venkataswamy Reddy, "A Census of Long-Stay Patients in Government Mental Hospitals in India," http://scribd.com/doc/70086163/A-Census-of-long-stay-patients-in-Government-Mental-Hospitals-In-India, Indian Journal of Psychiatry, 2001.

[9] PrabhaS.Chandra,V.Ravi, A.Desai, D.K. Subba krishna, "Anxiety and depression among hiv-infected heterosexuals-a report from India, "http://www.jpsychores.com/article/S0022-3999(98) 00028-2/abstract, Journal of Psychosomatic Research, February 1998.

[10] Das, Shankar and Leibowitz, George S.(2011) "Mental health needs of people living with HIV/AIDS in India: a literature review", AIDS Care, 23: 4, 417 — 425, First published on: 07 December 2010 (iFirst)

[11] T. Trojer, B.C.M. Fung, and P.C.K. Hung, "Service-Oriented Architecture for Privacy-Preserving Data Mashup," Proc. IEEE Seventh Int'l Conf. Web Services, pp. 767-774, July 2009.

[12] "India HIV&AIDS statistics," http://avert.org/india-hiv-aids-statistics.htm.