USER-CENTRIC PRIVACY PRESERVATION SOLUTION TO CONTROL THIRD PARTY ACCESS IN DIGITAL DATABASES

Muditha Tissera[1], Samantha Thelijjagoda[1], Jeevani Goonathilake[2]
[1]Sri Lanka Institute of Information Technology, Sri Lanka
[2]University of Colombo, Sri Lanka

ABSTRACT

The world is changing rapidly as technology advancements. Today, everything is e-enabled and human-linked information is digitally stored in digital databases, files and so on. Information privacy is a human right but, many privacy breaching incidents prove that, privacy has been threatened remarkably. Privacy is individualistic and dynamic in its nature. Many institutions, which collect human information, look at the privacy through the lens of a common pre-defined privacy policy or act. Such coarse-grain privacy preservation leads to violate individual's specific privacy requirements. Simply, user-centric privacy preservation is not guaranteed. The author attempts to solve this human critical problem by proposing a fine-grain privacy preservation solution. A unique conceptual framework is presented with a novel notion of Key Privacy Determinant Attributes (KPDA) Index. User's specific privacy perspectives will be captured through KPDA Index abstractly. However intelligent algorithms dynamically derive user sensitive database attributes to hide, when controlling third party access.

KEYWORDS: *User-centric Privacy, Digital databases, Third party access, Limited publishing.*

I.   INTRODUCTION

Rapid advances in Information and Telecommunication Technology are causing major behavioral changes in human life. People now live in an e-enabled digital world. As a long term consequence of this revolution, consciously or unconsciously someone's privacy would turn out to be revealed. For an instance a patient's medical data associated with his/her sensitive privacy information, captured by the medical practitioner (Primary user), may be accessed by pharmaceutical/medical health research organization (Secondary user/ Third Party user) or some employers (Secondary user/ Third Party user) may misuse such medical data for pre-employment suitability verifications or some Insurance companies (Secondary user/ Third Party user) may analyze such data to uncover some diseases of their clients, prior to set rates. All these domains keen to be aware of their client's undisclosed information [12]. Analyzing data and generating the intelligence for the well-being of human, is vital with no arguments. However, protecting the privacy of the Data Subject[1] (DS) should also be considered as a mandatory responsibility. Therefore, it is highly essential to control the usage of privacy sensitive data of a DS by Third party institutions.

The content of this article is organized as follows. After this introduction section, next defined what the Privacy is and how it differs from Security. Then, some key words relevant to the research such as User-centric privacy, Digital databases, Third-party access and Limited disclosure are comprehensively elaborated. Next, the main research problem/gap is declared. A new section is devoted to a broad review of literature relevant to the research. Conceptualization starts illustrating a framework that can be implemented to bridge the highlighted main research gap. It is explained with the help of Sub research objectives. A practical example is visualized throughout the conceptualization section for understandability.

---

[1]Owner of the data is referred as the Data Subject (DS) in this report

1.1 Define Privacy

Defining privacy is influenced by many cultural and sociological factors. The boundaries and content of what privacy means, vary among cultures and individuals, but hold common focus which control over human information sharing [21]. In many dictionaries privacy is defined as "someone's right to keep their personal matters and relationships secret or free from either public scrutiny or having the secrets or personal information shared". Kemp and Moore have highlighted the fact that "Privacy is difficult to define and has been challenged on legal and moral grounds; it is a cultural universal and has played an important role in the formation of Western liberal democracies". Further, they have provided an overview of the most important philosophical and legal issues related to privacy [2].

Preserving privacy of human is essential. That is why privacy rights are being established. Privacy rights such as the "right to be let alone, the option to limit the access others have to one's personal information", secrecy, or the option to conceal any information from others, control over others' use of information about oneself, the idea of personhood (there is something natural in being a human which requires humans to conceal some information) and the protection of intimate (interpersonal relationships)" have been identified by Solve and Daniel from many sources [33]. Having analyzed various privacy regulations and guidelines Agrawal, Kiernan et al. presented ten founding principles of Hippocratic database which demonstrates a detailed definition for privacy rights of a DS. This tenet includes "Purpose specification, Consent, Limited collection, Limited use, Limited disclosure, Limited retention, Accuracy, safety, Openness and Compliance" [1].

In the industry domain, majority appreciates the privacy preservation through the lens of a pre-defined privacy policy which is common to all DSs. Nevertheless, the worse part of this is, the implementation of such policies is a sole responsibility of the data collector (Primary users of data). Proper auditing is essential to avoid violating these privacy policies. Privacy Certificates like TRUSTe provides Data Privacy Management (DPM) solutions to ensure privacy compliance and build customer trust [39].

1.2 Privacy and Security

Security enforcements do not assure the privacy. Organizations extensively enforce security measures to protect their stored data from unauthorized access. Privacy is automatically preserved with security enforcements to a certain extent, but not guaranteed. As an example an employee who has legitimate access to the data can sell their clients' privacy sensitive information to third parties. It is not a security breach. But it is a privacy breach. Privacy and Security are two different aspects. Hence, it is vital to address privacy preservation separately in any digital database.

1.3 User-centric privacy

User-Centric Privacy Preservation is a forward-thinking that extends beyond the general policy driven privacy enforcements. Privacy is a human right. One should be able to decide what, when, to what extent and to whom his/her private data should be revealed.

According to Stone, Gardner et al, real privacy means "the ability of the individual to personally control information about one's self" [35]. From a recent study it was concluded that users are the key for next digital strategy. There is an imperative need in Europe to strengthen the user's role and perspective in the Information Society [7]. Privacy perspectives are very different from one DS to another. Not only that but also, it varies even within the DS throughout the life which brings the dynamic nature to human privacy. Based on the theory developed by social psychologist Irwin Altman in 1975; Leysia and Paul in 2003, outline a model of privacy as a dynamic, dialectic process [26].

Since, privacy is an individualistic human aspect; having a common single privacy policy for everyone, may not fulfill the real individual privacy needs of a DS. Common policy may increase the vulnerability of someone's individual privacy aspects to be breached.

Ex: Suppose, the first person is unhappy in disclosing his financial information while the second person is for marriage information. If their information has been collected by a financial institute and its policy prevents disclosing only financial information to third parties, the second person's privacy requirement may breach.

This implies the demand for a user-defined/user-centric privacy preservation. In fact, people would claim for their privacy rights based upon their own perspectives tomorrow than today.

1.4 Digital Databases

In this digital era, due to pervasiveness of internet (Especially with IoT, privacy in everyday life) and other electronic and telecommunication advancements, human linked information is captured and stored enormously. Few examples are highlighted below.

1) CCTV cameras monitor some areas and collect information about the people passing through those areas.
2) Some telephone conversations of DS may be recorded for the reasons such as security and improved quality of services.
3) Wireless Body Area Networks (WBAN) includes applications such as Endoscopic capsules, Heart rate monitors, Blood pressure monitors etc. track the human privacy sensitive information [3].
4) Smart location based applications track and save every moment of the DS.
5) Business applications capture its customers' personal and transaction information.
6) Research institutes capture survey data for various analyses.

All such collections may have privacy sensitive information. Those may be stored in digital databases, file systems or may be in the cloud. However, in this research, main focus is on digital databases of business applications which contain data values in tabular format (except multimedia files).

1.5 Third party access

DS is bound to provide information for organizations, which collect and process them as a Primary user to give various services to the DS. Examples could be hospitals, Department of Immigration and Emigration, Inland Revenue department, Police, Banks, Employers and Internet service providers etc. These primary users are given consent to use the information for the intended purpose of their business (Primary usage) only through the business application. Any kind of database access to sensitive information of DS which is explicit to the business application should be prohibited. When such data is reached to some other users' hand, lawfully or unlawfully from primary users, without getting the consent from the DS, it becomes a data/privacy breach. These other users can be called as Secondary users or Third party institutions such as research organizations, Information brokers and various other firms. Third parties keen to collect information about DS for many aspects such as generating intelligence, marketing (personalized advertisements etc.) and develop tailored systems (Ex. Adaptive systems) and many more [15].

Third party access can turn up in two means.

1) Authorized users intentionally disclosing information

Privacy Rights Clearing house (PRC) indicates that there are 4,128,666 records in their database for the period of years2011-2014, just only for the Insider(INSD) breach type which means "Someone with legitimate access intentionally breaches information" [29]. This Insider (INSD) breach type referrers to some authorized users such as employees of the organization intentionally disclosing privacy information of the DS to third party institutions.

2) Hackers/ Intruders unlawfully access privacy information

Various security enforcements prevail in handling such situations. But, data/privacy breaching is yet occurs.

Privacy Rights Clearing house (PRC) indicates that there are 173,977,329 records in their database for the period of years 2011-2014, just only for the Hacking or malware (HACK) breach type which means "Electronic entry by an outside party, malware and spyware" [29].

Though, the data is hacked, a good privacy model should not let the intruder to interpret privacy information.

1.6 Limited disclosure/ Limited publishing

As identified in Hippocratic databases, ten privacy principles assure the real privacy rights of the DS. Limited disclosure or Limited publishing is one out of those ten principles and defined as "The personal information stored in the database shall not be communicated outside the database or purposes other than those for which there is consent from the donor of the information." [1]. Policy driven privacy mechanisms implement limited publishing by having pre-defined set of database attributes identified by the organization and preserving them when disclosing data to third parties. True User-Centric privacy

enforcement means, individual DS preferred database attributes should be preserved. However, feasibility issues may come up when the database has large number of attributes; it is cumbersome for the DS to opt-out attributes which need to be preserved (Especially when giving the consent to disclose). Fulfilling this requirement is challenging but it is mandatory for user-centric and fine-grained privacy preservation. Therefore, a mechanism should be identified to perform this while not worrying the DS.

1.7 Main research problem

People do not willing to disclose their privacy sensitive data to third parties. However, organizations (primary users) who keep individuals' data, may decide to distribute such information to third parties fitting to the standard privacy rules, policies or acts. Every human has his/her own privacy perspectives. They may differ from one individual to another and even within one's self from time to time as well. This means, having a single policy that common to everyone, does not respect the real privacy needs of individuals. This leads to violate individual privacy perspectives. Some research approaches have demonstrated a certain degree of user-centric privacy. However, a clear knowledge gap prevails to give a guarantee to the DS in fulfilling the user-centric privacy preservation in digital databases.

1.8 Main Research Objective

The main objective of this research is to provide a conceptual framework for a comprehensive design solution to preserve privacy in digital databases to control over the accessibility by third party institutions. This research is unique because, the suggested access control mechanism is based upon the limited disclosure of data by respecting specific privacy perspectives of individual DS. (User-Centric)

II. RELATED WORK

Many technological framework types have been proposed for Privacy preservation in digital databases. Limited Publishing, Data Perturbation and Data Encryption are the three frequently used types in many research work [11].

Hippocratic database concept suggested by Agrawal and the research team in 2002, has become the unified architecture and starting point for various other research extensions. They believe that the future databases must be responsible for privacy preservation as its' founding principle. The value of this research is mainly fall on the identified ten principles based upon the Hippocratic auth. They introduced a strawman design of a Hippocratic database system. In this design, they use "purpose of the data usage" and the "recipient" as the central concept around which the privacy protection is built. Limited publishing and Limited retention are in-built with their design. The main flip side of this concept is, it requires the set of predefined external recipients, when the information is collected to the database [1]. LeFevre and the research team proposed another extension to this limiting disclosure in Hippocratic databases. In this model, Privacy policies are created for user groups and those will be stored as privacy meta-data in the database. Each query should be attached with the purpose and the recipient. SQL queries are modified using a query semantic model and remove the prohibited data from the query result set based on the purpose. Since, Hippocratic database has become the base for this model, main drawback of so called original concept has been inherited to this. Another drawback which hiders the performance of this model is, it requires to create views for each Purpose-Recipient pair [17]. Various other research works exist based on this Hippocratic database concept with many different variations. Proposing some hierarchical approach to Hippocratic databases [22], determines the minimal set of data which needs to fulfill the purpose. Then, collection of data would be limited accordingly (Limited collection). Realizing privacy-preserving features in Hippocratic databases [43] provides extended features which include support of multiple policy versions, retention time, generalization hierarchies, and multiple SQL operations. Hippocratic database model which is designed to relational databases extended into the Hippocratic XML database model [16]. Another extension to Hippocratic database can be found in large scale video databases to preserve video content privacy [13]. In this research, privacy-sensitive human objects have been successfully filtered out automatically from the video content. Hippocratic PostgreSQL has the novel feature of supplementing it with both k-anonymity and generalization hierarchies into Hippocratic DBMS implementation [25]. Optimal Privacy-aware path in Hippocratic databases is identified using purpose directed graph (PDG) [18]. Another study has been conducted for privacy preserving feature for education industry in Hippocratic databases by Bedi &

Thengade [6] . All above Hippocratic database related research work requires restructure of the DBMS. They all have attempted to preserve at least one or more privacy principles (out of 10) uniquely in their work.

A novel privacy preservation method based on Data Perturbation approach has been suggested by Kun Guo and Qishan Zhang. The key theory behind this method is the GM(1,1) model based on Grey Systems theory. GM(1,1) model aims at discovering grey information in the data and whitening it to reduce the uncertainty. In this way the hidden relations can be uncovered. On the contrary, the goal of privacy preservation is to obfuscate the certain data to protect persons' privacy. From the view of grey system theory, the uncertainty of the data can be increased by introducing grey information, i.e. graying the white information. So the authors have reversely applied the GM(1,1) model to perturb the data [11].

Biometric data (Ex: finger prints, iris) is commonly used nowadays in security enforcements because of their higher accuracy for personal recognition. The flip side of this is such data is linked with an individual and if compromised, it leads to privacy violations. Fully Homomorphic Encryption (FHE) has been used [38] as a solution for the privacy issue because of its ability to perform computations in the encrypted domain. Data Encryption approach is the base for this research.

Privacy-preserving in distributed event-based systems specifically for Streaming Databases has been approached using cryptographic techniques by Oliveira et al. They examined the performance of several privacy preserving data correlation techniques using symmetric encryption [32].

Anonymization is another technique that can be used in privacy preservation. The concept of k-anonymity was formulated by Latanya Sweeney to solve the problem: "Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful." In this approach, privacy is enforced through the use of Generalization and Suppression. Further, in addition to the main k-anonymity protection model author has explored privacy related attacks and provided ways in which these attacks can be controlled [37].

Record linkage and attribute linkage attacks are the major attacks against privacy. Mahesh & Meyyappan [10] highlights that K-anonymity method preserves the privacy against record linkage attack alone. It fails to address attribute linkage attack. l-diversity method overcomes the drawback of k-anonymity method. But it fails to address identity disclosure attack and attribute disclosure attack in some exceptional cases. t-closeness method preserves the privacy against attribute linkage attack but not identity disclosure attack. But its computational complexity is large. Therefore, authors introduce a new method which overcomes all such issues. Their approach is to generalize quasi identifier by setting range values and record elimination.

In anonymization, the anonymity of individuals should be protected. Though, Data holders remove or encrypt explicit identifiers such as names, addresses and phone numbers, other distinctive identifiers are combined uniquely and linked to publicly available information to re-identify individuals. Samarati & Sweene attempted disclosing entity information such that the released table cannot be reliably linked to external tables [31].

The k-anonymity algorithm has been used by Eltabakh et al as a relational operator in DBMS [8]. That interacts with other query operators and applies the privacy requirements when query requests are being processed. This model provides limited disclosure and limited retention also with anonymized views. Authors have selected the Hippocratic database as the base for this experimental research.

Assuming that different attributes in the database may need different privacy levels of protection, personalized privacy preserving method using "Randomized Response Technique." has been introduced by Sun et al [36]. This particular solution is appropriate for market basket which data is in binary format. However, attribute level privacy doesn't aim for user-centric privacy while row level serves the purpose. In another article, authors have outlined an approach to valid statistical analysis of distributed data that does not require actually integrating the data. Instead, it is based on anonymized sharing of database-specific sufficient statistics in such a way that no database owner can separately identify the individual contributions of the other owners [14].

In many data mining approaches, information loss is visible when privacy preserving. Menon & Sarkar suggest an integer programming formulation which minimizes the number of non-sensitive itemsets lost while hiding sensitive itemsets. For this, authors use Frequent ItemsetHiding (FIH) algorithm in a modified way through sanitization problem (identify how the support for sensitive itemset can be

eliminated from a tuple by removing the fewest number of items from it) This problem arises when databases are shared between firms. Approach is more applicable when the receiver shares their mining threshold with the owner of the data. This approach preserves privacy. However, when sensitive data finds, entire record is undisclosed. Expected user-centric privacy cannot be seen here [23].

A comprehensive survey on privacy preserving data mining has been done by Wang Jet al. in 2009. They have highlighted many technological approaches and their usages by different researches with their merits and shortcomings. Highlighted technologies include K-Anonymity, Perturbation approach, Cryptographic techniques, Randomized Response technique and the Condensation Approach [40]. A similar survey has been done in 2016 also but in addition to the techniques used by the previous authors, they have studied two new approaches, Blocking based technique and a Hybrid technique which uses Randomization and Generalization. Having compared different approaches based on performance, utility, cost, complexity and tolerance against data mining algorithm, they have concluded that all methods behave in a different way depending on the type of data or type of application domain. But still Cryptography and Random Data Perturbation methods perform better than other methods. [45]

Lu, Zhu et al, have researched on Privacy-Preserving Computing in Big Data Era. First they have identified privacy requirements in big data from different angles such as Privacy requirements in big data collection, big data storage and big data processing. While discussing different existing privacy preserving techniques such as privacy preserving aggregation, Operations over encrypted data and De-identification, they have proposed an efficient protocol called privacy-preserving cosine similarity computing protocol as a solution for privacy requirements of data mining in the big data era [19].

Internet privacy is also has become a hot topic due to pervasiveness of internet and social networks. Collaborative filtering technique gives good recommendations using internet data to many of peoples' daily activities. However, it threatens to privacy. Accuracy and Privacy becomes a tradeoff. (To preserve privacy they tend to use inaccurate data) The authors of the paper on "Privacy Preserving Collaborative Filtering"[28], presents a solution to effective Collaborative Filtering with Privacy by balancing both privacy and accuracy. Randomized perturbation technique is used to disturbed user's personal information including his/her voting rates (likes or dislikes etc.). This perturbation prevents server to learn the user but allows perform collaborate filtering using perturbed data.

Ma, Meng & Wang have proposed a novel privacy threat in the internet called "Privacy Inference Attack via Search Engines". Attacker can gather user's scattered information in the internet using search engines and mine some privacy information. In this research approach, it constructs a user information correlation (UICA) graph to model the association between user information returned by search engines. Then it assigns some probability value to vertices and use greedy algorithm to identify the maximum probability path which helps to reveal the level of vulnerability for the attack [21].

In another research, privacy preserving data publishing mechanisms were evaluated in detail. In one of their exploring methodologies, an approach was identified and discussed about data publishing in social networks. It explained that the social network can be modeled as a simple, undirected graph $G = (V, E)$, mapping nodes correspond to entities and edges represent connections between entities. Each entity has a unique name. The goal of this research is to publishing data on the internet while preserving privacy. Removing information of entities (nodes) using naive de-identification approach while preserving the topology of the graph (different relationships) was the logic behind these approaches [44].

There are some other privacy related researches which are not directly related to privacy in digital databases. But their conceptual frameworks are appreciated.

Trust base privacy preservation is also another method used in privacy preservation. Ruotsalainen and the team in 2013 developed a privacy management architecture that helps the DS to create and dynamically manage the network and to maintain information privacy. Trust models such as belief, organizational trust, dispositional trust, recommended trust and direct trust are used in privacy preservation. The architecture should provide to the DS reliable trust information about systems and assist in the formulation of privacy policies [30].

An agent base privacy preservation method has been introduced by Stamp and Lee. This model employs a collection of software agents which Privacy-related decisions are made on behalf of the user [34].

In addition to the above frameworks, there are some other well-known privacy related theories such as Privacy regulation theory that was developed by social psychologist Irwin Altman in 1975 [5]. This theory explains why people sometimes prefer staying alone but at other times like get involved in social

interactions. Further, Petronio's communication privacy management (CPM) theory is an important extension of Altman's theory that particularly suited for the study of social networking [27].
The different privacy concepts that are in academic literature have been brought in to a single discussion and distinguished them from one another with advantages and disadvantages of such definitions by Thomas Allmer. The author highlights that there was a gap in the existing literature for a definition to privacy. Based on the thoughts of Lyon [20] and Fuchs [9], Allmer says social theories can be used to define the privacy. "Social theories deal either with social structures, or/and with social actors." Findings allow distinguishing privacy definition by three aspects such as Structuralistic, Individualistic, and Integrative. The article proves using studies of other authors including Westin [41] [42], and Moor [24], "Individualistic approaches of defining privacy focus on the individual and understand privacy as control over information about oneself" [4]. Therefore, this Individualistic aspect is much closer to the User-Centric Privacy enforcement that is discussed in our study.

III.   CONCEPTUALIZATION

In order to achieve the main objective of this research, the following conceptual framework is proposed as a solution design. Figure 1, summarizes the proposed concept as a high-level architecture diagram.
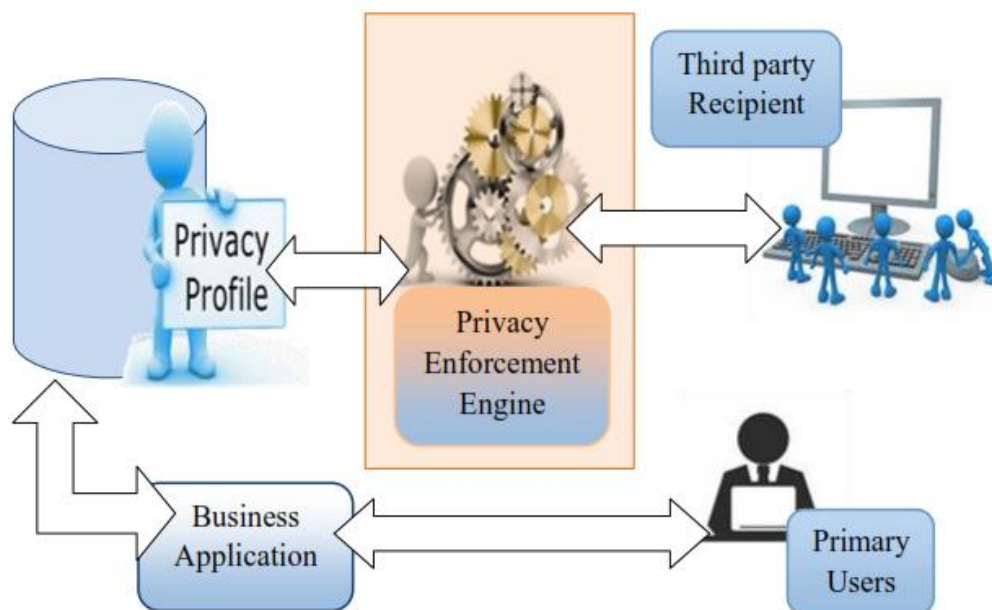

Figure 1: High-level Conceptual Framework

To enforce user-centric, fine-grained privacy preservation, it is mandatory to identify specific privacy attributes/concerns of individual DS and generate a Privacy Profile for each DS. Privacy Enforcement Engine is one of the key components of this novel conceptual model. It should consist of many different intelligent algorithms to virtually wrap the database to third party institutions based upon the identified individual Privacy Profiles of DSs. Due to that, third parties should never have access to privacy sensitive data of the DS. User-centric privacy would be preserved. In this suggested conceptual framework; primary users will not be affected with limited disclosure through User Privacy Profiles when they interact through the business application.
The above high-level architecture can be further elaborated into much detail via the Detail-level Conceptual Framework depicted in Figure 2. It identifies different work components of the research as Sub Objectives to be achieved.
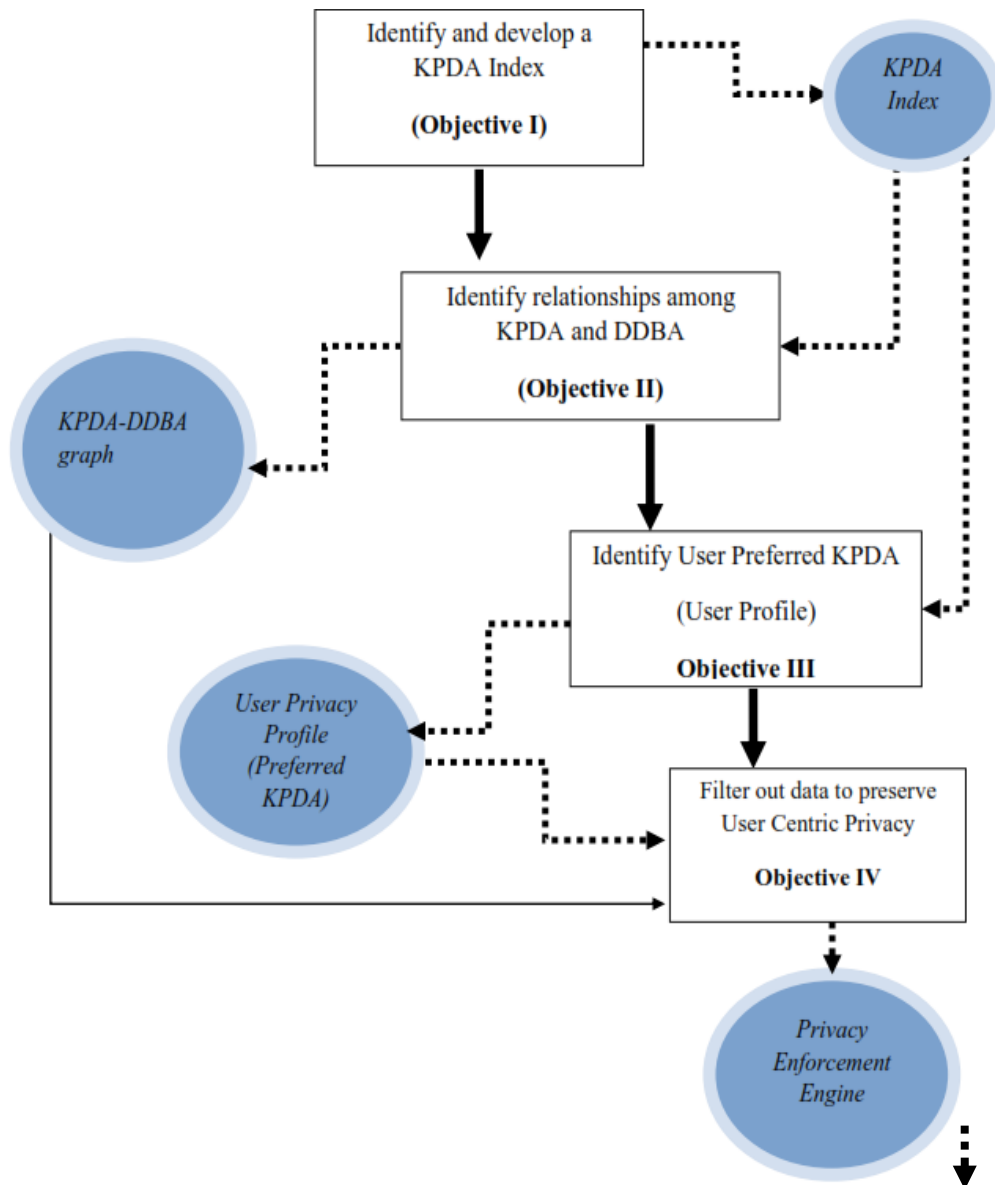
Figure 2: Detail-level Conceptual Framework

IV.   SUB RESEARCH OBJECTIVES

The intention of this research is to enforce user-centric privacy in digital databases. Business application specific databases are huge in size as in their nature. In such databases, the privacy sensitive information of the DS may have been spread among thousands of fields in hundreds of tables. Due to this gigantic nature, it is not feasible asking the user to specify which database attributes need or need not to disclose. However, that is crucial in the real user-centric privacy preservation. In other words, the individual Privacy Profile of the DS should be recognized. Even though, prompting thousand-odd attributes one by to DS not feasible, the author presents a novel notion of Key Privacy Determinant Attribute Index (KPDA Index) in this conceptual model that will overcome this critical issue. For better understanding, each sub objective is visualized below in a graphical mode. Table I below is the Legend used in the visualization.

Table 1: Legend

KPDA	[green square]
DDBA	[blue circle]
User identified sensitive KPDA of DS (in UPP)	[green square with red border]
System identified sensitive DDBA of DS	[blue circle with red border]

Sub Research Objective 1 - To develop a KPDA Index.

Investigate Key Privacy Determinant Attributes (KPDA) that would cover the entire human privacy spectrum and develop a KPDA index. Through a preliminary survey and from the existing literature, it is necessary to find out this privacy determinant attributes. As depicted in Table 2, KPDA can be categorized as Main and Sub attributes. KPDA index will be primarily used for two important aspects in this research. Firstly, it will be used to identify the User Privacy Profiles. Secondly, it will be used to generate the KPDA-DDBA graph model. (These will be further discussed under the next sub sections) This initial attempt will result the first version of such KPDA Index. Researching and improving this first version to a next upgraded version will be kept open as a future work.

Table 2 - KPDA Index

Key Privacy Determinant Attributes (KPDA) Index

KPDA Main	KPDA Sub	Public Information	Less Sensitive Information	Sensitive Information	Very Sensitive Information	Critically Sensitive Information
Medical	Critical illnesses					
	Medicine					
	Surgeries					
	Allergies					
	Mental Illnesses					
Financial	Remuneration					
	Savings					
	Assets					
	Shares					
	Loans					
	Insurance					
	Tax payments					
	Credit Card Numbers					
Personal Contact Information	Name					
	Phone Numbers					
	Email addresses					
	Social Identification Numbers					
	Address					
Employment	Designation					
	Name of the Organization					
	Address					
	Office Email					

Sub Research Objective 2 - To develop a KPDA-DDBA Graph Model.

As the first step of this Sub research objective, identified KPDA (All sub attributes) should be represented as nodes in a graph as visualized in Figure 3. (In the following example, some randomly picked KPDA has been used for easy understanding.)
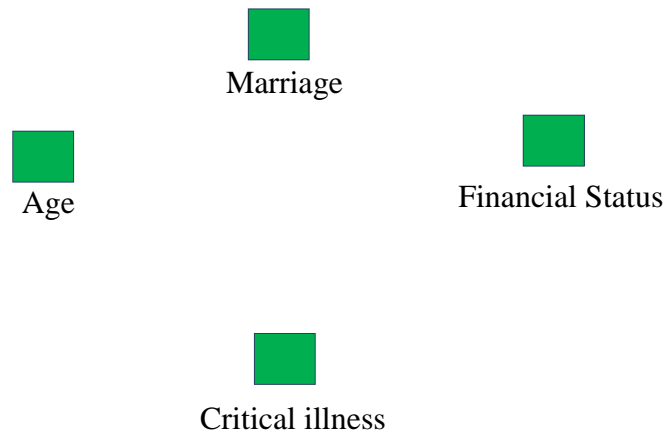
Figure 3: - KPDA set as nodes forming a graph.

As the second step, business application specific digital database attributes (DDBA) should also be represented as nodes in the same graph. When all such nodes are set, it is essential to derive an intelligent classification mechanism to identify the relationships among KPDA and Digital Database Attributes (DDBA) of the database by considering their closeness, relevance etc with respect to privacy and model such relationships as edges in the graph.

Ex: "Marriage" KPDA can relate to DDBA attributes such as "Marital status", "No of kids", "Maiden name", "C-section Birth" while it does not relate to "Blood group" attribute in a medical / health care database.
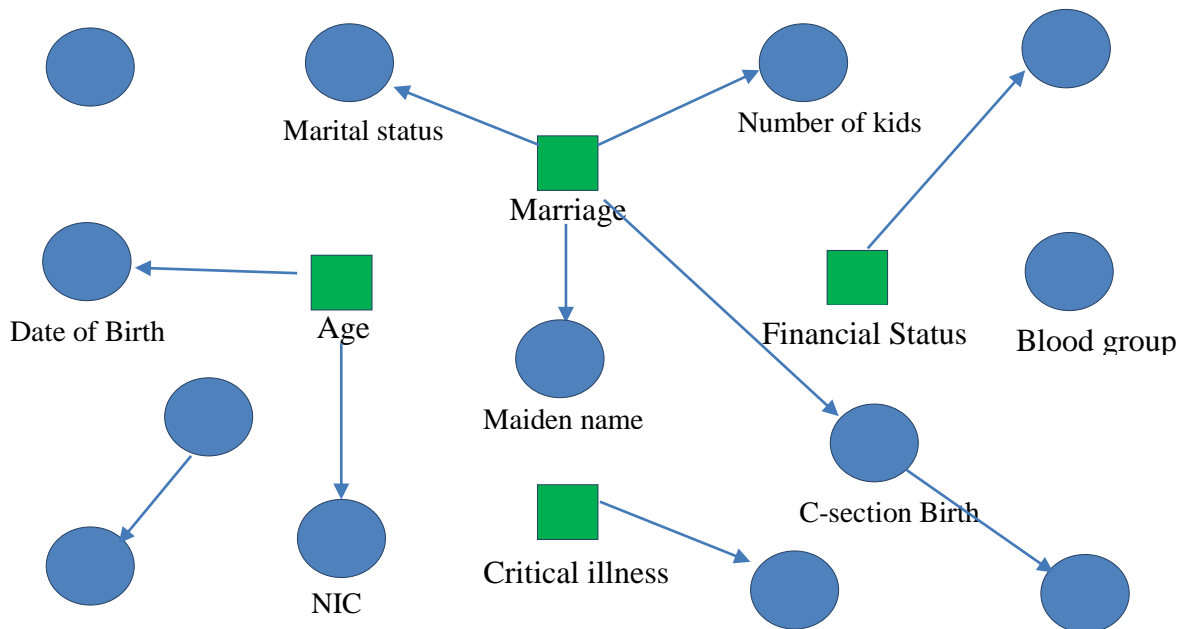

Figure 4: Step 2 – KPDA-DDBA Graph model.

This mapping would finally generate a KPDA-DDBA graph model, as visualized in Figure 4. However, the biggest challenge in this step is how to accurately identify these relationships (or edges of the graph). If the application/database designer is asked to perform this manually, a good accuracy can be expected. This may feasible either the size of the database is fairly small or mapping is done at the time of database schema design. However, in some existing business applications (specially in legacy systems), database may be very complex and huge in size. There may be thousands of attributes in hundreds of tables. In such cases, identifying relationships manually would be a cumbersome and time consuming task.

Therefore, it has to be automated and a solid technological solution should be proposed. A proper intelligent classification mechanism should be identified to automate this.

Sub Research Objective 3 - To generate User Privacy Profiles (UPP)
Individual DS can be given the KPDA index and request to mark his/her own sensitive attributes in the KPDA Index. Number of attributes in a KPDA index (less than hundred) is much smaller when it compares with the number of attributes in a Database (more than thousand). Hence, the DS should be able to mark his/her sensitive KPDA from the KPDA Index. Such selection represents own privacy profile of the DS. This UPP will broadly envisage the individual's specific privacy perspectives.

Table 3 –User Privacy Profile

User Privacy Profile (UPP) on KPDA Index

KPDA Main	KPDA Sub	Public Information	Less Sensitive Information	Sensitive Information	Very Sensitive Information	Critically Sensitive Information
Medical	Critical illnesses				√	
	Medicine			√		
	Surgeries			√		
	Allergies					√
	Mental Illnesses					
Financial	Remuneration				√	
	Savings				√	
	Assets				√	
	Shares				√	
	Loans			√		
	Insurance		√			
	Tax payments	√				
	Credit Card Numbers					
Personal Contact Information	Name	√				
	Phone Numbers			√		
	Email addresses			√		
	Social Identification Numbers			√		
	Address			√		
Employment	Designation		√			
	Name of the Organization	√				
	Address	√				
	Office Email			√		

Different DS may have different degrees of sensitiveness for different KPDAs. This degree of sensitiveness can be successfully captured using a likert scale such as Public, Less sensitive, Sensitive, Very sensitive or Critically sensitive etc. (Though, the real implementation should have all such degrees of sensitiveness depicted in Table 3, in the example visualized here, doesn't represent all such different degrees. Instead it assumes that user is given only two options, public or sensitive for easy understanding and visualization in the example)
DS transacts with business organizations and interacts with their business applications. During such interaction, (Ex. As a part of the client registration process etc.) the KPDA index can be prompted to DS to mark their privacy preferences. (Example is depicted in Table 3)

Sub Research Objective 4 - To design a Privacy Enforcement Engine
This should be an access control mechanism to the database, which should preserve user-centric privacy through User Privacy Profiles.
As visualized in Figure 5, User privacy profile can be virtually mapped on top of the identified KPDA-DDBA graph.
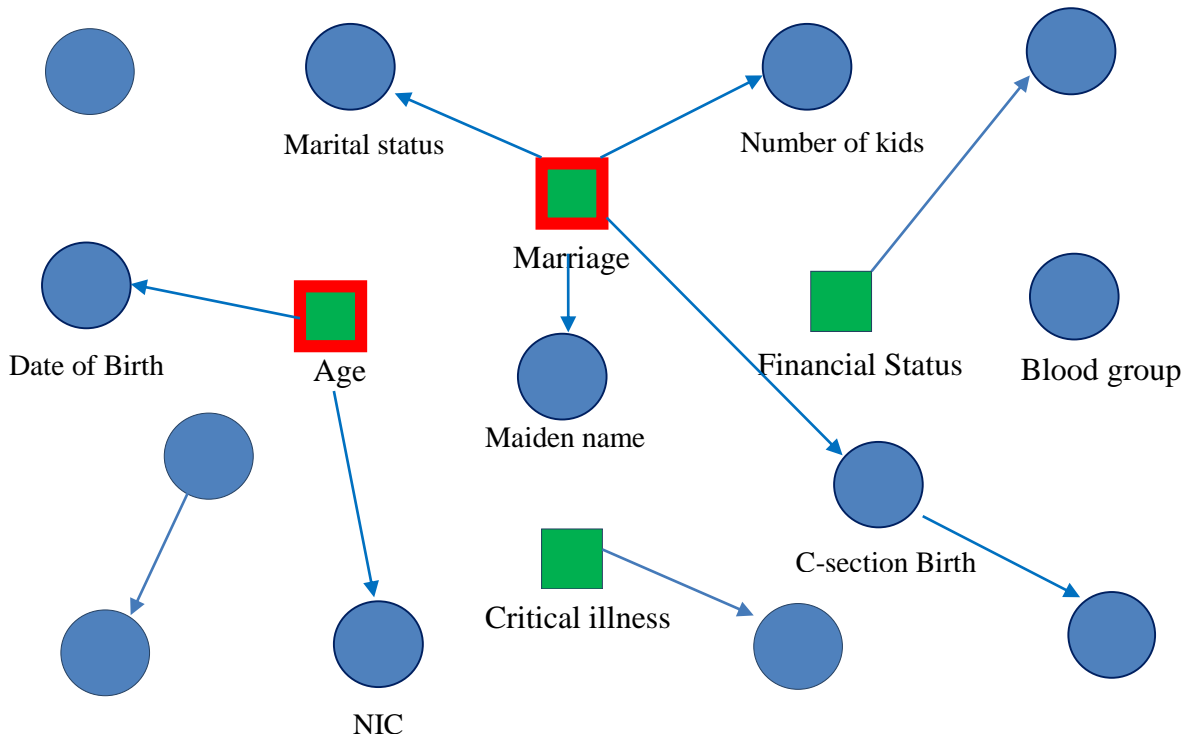
Figure 5: User sensitive KPDA (UPP) virtually mapped on top of KPDA Index

DDBAs that are connected to user identified sensitive KPDAs (in UPP) (directly or through another connected node) can be considered as real privacy sensitive DDBAs of DS. Hence, using this graph model, it is possible to automatically identify DS specific sensitive DDBA without any human intervention. For more clarification, refer the example below and the Figure 6.

Ex: Assume that the UPP has Marriage and Age marked as sensitive KPDA.

Marriage [√]   Age [√]   Financial Status [ ]   Critical Illness [ ]

With the support of the developed KPDA-DDBA Graph, following DDBA would be automatically selected (by Privacy Enforcement Engine) as privacy sensitive DDBA of DS.

  i.   Marital status
 ii.   No of kids
iii.   Maiden name
 iv.   C-section Birth
  v.   Date of Birth
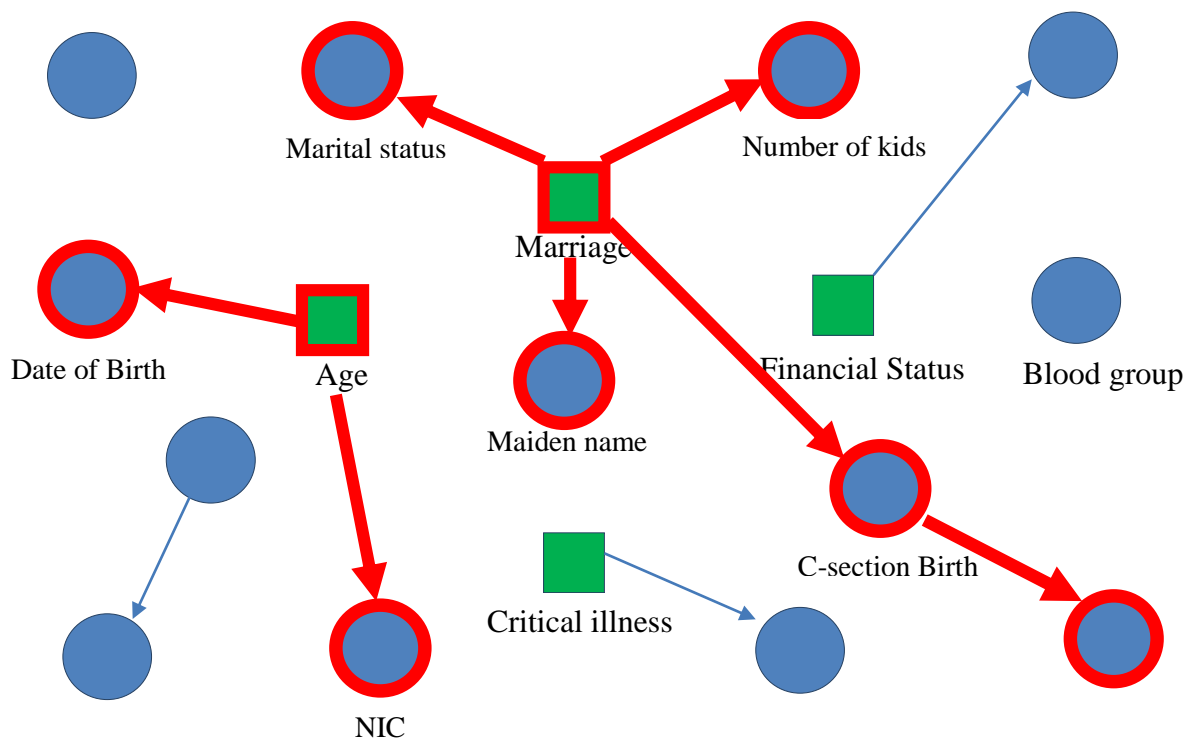 vi.   NIC (Date of birth can be derived from this)

Figure 6:   Step 4 - Apply User privacy profile and filter data.

With this novel design approach, DS will not be asked to manually opt out attributes that he/she needs to preserve/undisclosed, one by one from the large number of attributes of the database. But his/her real privacy requirements would be preserved automatically. Irrespective of the number of Database Attributes (DDBA) in the digital database, Individual's sensitive attributes are automatically identified. Though, the UPP shows a kind of abstract nature, with this conceptual design it is expected to figure out near real individual privacy perspectives in to a considerable fine-grain level. That helps to achieve user-centric privacy preservation.

V.   CONCLUSION

5.1 Significance of the research

Due to fast growing technology advancements, digitization of information is rapidly increasing. As an adverse consequence of this digitization, information privacy has been threatened enormously. Information privacy is a human right. Rather than just anticipating into an ethical behavior from the data collecting organizations to comply with a common privacy policy, a technological underpinning on behalf of the Data Subject should be provided. Research attention prevails for this purpose and it has been critically reviewed in the Literature Survey section. However, individual users may not be happy because, their individual privacy perspectives are not respected amply. User-centric privacy is a kind of forward thinking that we must prepare as early as possible. In this research, a significant novel approach is proposed to fill the prevailing gap in the need of a user-centric fine-grain privacy preservation solution in digital databases. Hence, it directly influences the preservation of an important human right with respect to information privacy.

5.2 Expected contributions to the body of knowledge

This research primarily attempts to propose a unique conceptual framework to preserve user-centric privacy in digital databases.

In addition to the so called main contribution, this research provides few other contributions to the body of knowledge. These include, introducing a useful notion of Key Privacy Determinant Attributes (KPDA) Index that can be utilized for other privacy related research as well, development of an innovative mechanism to simulate human thinking, when automating the identification of relationships among KPDA and DDBA to produce the graph model, and finally the development of intelligent query processing algorithms that enforce user-centric privacy preserved access to digital databases while searching through a graph to filter out sensitive data.

REFERENCES

[1]. Agrawal, R., Kiernan, J., Srikant, R., &Xu, Y. (2002). Hippocratic databases. In Proceedings of the 28th international conference on Very Large Data Bases (pp. 143–154).

[2]. Adam R. K., Moore D., (2007),Privacy, Library Hi Tech, 25(1), 58 – 78

[3]. Al Ameen, M., Liu, J., & Kwak, K. (2012).Security and Privacy Issues in Wireless Sensor Networks for Healthcare Applications. Journal of Medical Systems, 36(1), 93–101.

[4]. Allmer T., (2011), A critical contribution to theoretical foundations of privacy studies, Journal of Information, Communication and Ethics in Society, 9 (2), 83 – 101.

[5]. Altman, I (1975).The environment and social behavior. Monterey, CA: Brooks/Cole.

[6]. Bedi, R. K., & Thengade, A. M. (2010). Purpose-based access control exploits by HDB. International Journal of Computer Applications, 1(6), 93–97.

[7]. Deloitte (2010), Background Document in Support of the Digital Agenda for Europe. Final Report to the European Commission, Deloitte, Brussels.

[8]. Eltabakh, M. Y., Padma, J., Silva, Y. N., He, P., Aref, W. G., & Bertino, E. (2012). Query processing with K-anonymity. International Journal of Data Engineering (IJDE), 3(2), 48–65.

[9]. Fuchs, C. (2008), Internet and Society: Social Theory in the Information Age, Routledge, New York, NY

[10]. Mahesh, R., Meyyappan, T. (2013). Anonymization Technique through Record Elimination to Preserve Privacy of Published Data. Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, (pp 21-22).

[11]. Guo, K. & Zhang, Q., (2012), Privacy preserving method based on GM(1,1) and its application to clustering, Grey Systems: Theory and Application, 2, 157 – 165.

[12]. Levitan D. (2015, May 28). KáriStefánsson Says Medical Privacy Is Overrated. [Online] Available:http://spectrum.ieee.org/biomedical/ethics/qa-kri-stefnsson-says-medical-privacy-is-overrated

[13]. Peng J, Babaguchi, N., HangzaiLuo, YuliGao, &Jianping Fan. (2010). Constructing Distributed Hippocratic Video Databases for Privacy-Preserving Online Patient Training and Counseling. IEEE Transactions on Information Technology in Biomedicine, 14(4), 1014–1026.

[14]. Karr A. F., Fulp W. J., Vera F., Young S.S., Lin X.& Reiter P. J., (2007), Secure, Privacy-Preserving Analysis of Distributed Databases, Technometrics, 49,3, 335-345

[15]. Kolter, J. P. (2010). User-centric Privacy: A Usable and Provider-independent Privacy Infrastructure (Vol. 41). BoD–Books on Demand.

[16]. Lee, J.-G., Whang, K.-Y., Han, W., & Song, I. (2006). Hippocratic XML databases: a model and an access control mechanism. COMPUTER SYSTEMS SCIENCE AND ENGINEERING, 21(6), 395

[17]. LeFevre, K., Agrawal, R., Ercegovac, V., Ramakrishnan, R., Xu, Y., & DeWitt, D. (2004). Limiting disclosure in hippocratic databases. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 108–119).

[18]. Li, M., Sun, X., Wang, H., & Zhang, Y. (2009).Optimal privacy-aware path in hippocratic databases.In Database Systems for Advanced Applications (pp. 441–455).

[19]. Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014).Toward efficient and privacy-preserving computing in big data era. Network, IEEE, 28(4), 46–50.

[20]. Lyon, D. (1994), The Electronic Eye: The Rise of Surveillance Society, University of Minnesota Press, Minneapolis, MN.

[21]. Ma, R., Meng, X., & Wang, Z. (2012).Preserving privacy on the searchable internet. International Journal of Web Information Systems, 8(3), 322–344.

[22]. Massacci, F., Mylopoulos, J., & Zannone, N. (2005).Minimal disclosure in hierarchical hippocratic databases with delegation. In Computer Security–ESORICS (pp. 438–454)

[23]. Menon, S., & Sarkar, S. (2007). Minimizing information loss and preserving privacy. Management Science, 53(1), 101–116.

[24]. Moor, J. (1997), Towards a theory of privacy in the information age, Computers and Society, 27(3), 27-32.

[25]. Padma, J., Silva, Y. N., Arshad, M. U., &Aref, W. G. (2009). Hippocratic PostgreSQL. In Data Engineering, ICDE'09. IEEE 25th International Conference on (pp. 1555–1558).

[26]. Palen L, Dourish P (2003). Unpacking 'Privacy' for a Networked World. Proceedings of ACM Conf. Human Factors in Computing Systems CHI'03 (Ft. Lauderdale, FL), 129–136.

[27]. Petronio, S. (2002). Boundaries of Privacy: Dialectics of Disclosure. Albany, NY: SUNY Press

[28]. Polat H. & Du W., (2005), Privacy-Preserving Collaborative Filtering, International Journal of Electronic Commerce, 9(, No. 4 (Summer, 2005), pp. 9-35

[29]. Privacy Rights Clearing house PRP.(2016, August 10).Chronology of Data Breaches, [Online] Available: https://www.privacyrights.org/data-breach/new

[30]. Ruotsalainen P, Blobel B., Seppälä A. and Nykanen A., (2013), Trust Information-Based Privacy Architecture for Ubiquitous Health, JMIR Mhealth Uhealth, 1(2): e23.

[31]. Samarati, P., & Sweeney, L. (1998).Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

[32]. Santana de Oliveira, A., Kerschbaum, F., Lim, H. W., & Yu, S.-Y.(2012). Privacy-Preserving Techniques and System for Streaming Databases (pp. 728–733).

[33]. Solove& Daniel, J. (2008). Understanding Privacy. Cambridge, Mass.: Harvard University Press.

[34]. Stamp M, Lee H., (2008), An agent-based privacy enhancing model, Information Management & Computer Security

[35]. Stone E. F., Gardner D. G., Gueutal H. G., and McClure S., 3 (1983), A Field Experiment Comparing Information-Privacy Values, Beliefs, and Attitudes Across Several Types of Organizations, Journal of Applied Psychology 68, 459–468.

[36]. Sun, C., Fu, Y., Zhou, J., &Gao, H. (2014). Personalized Privacy-Preserving Frequent Itemset Mining Using Randomized Response. The Scientific World Journal, 2014, 1–10.

[37]. Sweeney L., (2002),k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 ,557-570.

[38]. Torres W. A. A., Bhattacharjee N., Srinivasan B., (2015), Privacy-preserving biometrics authentication systems using fully homomorphic encryption, International Journal of Pervasive Computing and Communications, 11, 151 – 168

[39]. TRUSTe Data Privacy Management Solutions. (2016, December 01). Ensure Privacy Compliance and Build. [Online] Available: https://www.truste.com/

[40]. Wang, J., Luo, Y., Zhao, Y., & Le, J. (2009).A Survey on Privacy Preserving Data Mining (pp. 111–114).

[41]. Westin, A. (1967), Privacy and Freedom, Atheneum, New York, NY.

[42]. Westin, A. (2003), Social and political dimensions of privacy, Journal of Social Issues, 59(2),(pp. 431-453)

[43]. Silva Y.N., and Aref, W. G., "Realizing Privacy-Preserving Features in Hippocratic Databases" (2006), Computer Science Technical Reports. Paper 1665.

[44]. YongbinY., Yang J., Zhang J, Lan S and Zhang J, (2011), Evolution of Privacy-Preserving Data Publishing, Anti-Counterfeiting, Security and Identification (ASID), 2011 IEEE International (pp. 34-37)

[45]. Prasanthi K., (2016), A review on Privacy Preserving Data Mining techniques, International Journal of Advanced Research in Computer Science and Software Engineering, 6(3), (pp. 35-40)

AUTHORS' BIOGRAPHIES

Muditha Tissera, Having played a lead development role in the software development industry for more than 15 years, Muditha Tissera changed her career into academia in 2013 as a Senior Lecturer at Sri Lanka Institute of Information Technology. After graduating with B.Sc (Management Information Systems) second class - Upper division honors from the National University of Ireland, she then obtained her MSc (Computer Science) with a Distinction pass from University of Colombo, School of Computing, Sri Lanka in 2001. Presently, she is a PhD student of Faculty of Graduate Studies, University of Colombo and her research interests include Privacy in Digital Databases, Data Science and Product line Software Engineering. She enjoys lecturing to undergraduate and Master degree students while supervising their research projects. Majority of her research collaborates with industry organizations to achieve implementation reality.

Samantha Thelijjagoda is a senior lecturer (Higher Grade) in Information Systems Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka. He is currently serving as the Acting Dean, Faculty of Graduate Studies and Research.
He received his first degree in Statistics with first class honors from University of Sri Jayewardenepura, Master of Engineering in Computer Systems Engineering and PhD in Information Systems Engineering from University of Gifu, Japan. His research interests are Computational models of human language processing (NLP), Human language technology (HLT) such as Machine Translations, Information Extraction etc. and Digital

linguistics (Corpus Linguistics) which are associated with the area of Computational Linguistics.

He is an active member of Computer Society of Sri Lanka and currently the student counselor of its executive council. He is the country representative of Technical Committee 8 (TC-8: Information Systems) of International Federation of Information Processing (IFIP). He is also the country in-charge for Skills Certifications of IT professionals in Sri Lanka, which is awarded by Australian Computer Society.

Jeevani Goonetillake is a senior lecturer attached to the University of Colombo School of Computing. She joined the University of Colombo in 1996 after graduating with B.Sc (Computer Science) first class honors from the same university. She then obtained her MSc (Data Engineering) from Keele University, UK in 1998 and PhD from University of Wales, UK in 2004 respectively. Her PhD research was based on exploring the issues relevant to collaborative engineering design environment and thus her PhD thesis is on "Managing Evolving Constraints in Collaborative Engineering Design Environment". Her research interests include database security and forensics, wireless and sensor network databases, NoSQL databases, social life networking and crowd sourcing. Currently she is supervising several MPhil and undergraduate research projects in these areas.