Big Data Privacy Risk

Connecting Many Large Data Sets

Star Ying
Lead Data Scientist,
U.S. Department of Commerce,
1401 Constitution Ave NW,
Washington, DC 20230.
sying@doc.gov

Tyrone Grandison
Chief Information Officer,
Institute for Health Metrics and Evaluation,
2301 Fifth Ave,
Seattle, WA 98121.
tgrand@uw.edu

***Abstract* — Data is the fuel, the glue and the product of online collaboration. Big Data is the driving force behind collaborative computing and is enabling and facilitating the next wave of innovation. Unfortunately, privacy is one of the core weaknesses of the entire ecosystem. The prevailing wisdom is that sensitive data can be protected in Big Data sets. In this paper, we decompose the problem space and mathematically discuss the implications for privacy when one connects the many, large data sets that comprise a Big Data collection.**

Keywords — Big Data; Privacy; Collaboration; Risk

I. Introduction

The term *Big Data* has gone through peaks and valleys of interest over the last few years. A cursory check of Google Trends on the search interest in the term shows that there has been an overall increase in activity since October 2011, with maximum interest being reached around February 2016 (Figure 1).

Fig. 1: Search Popularity of Big Data.
Source: Google Trends (Oct 6th, 2016).

However, the term Big Data has been floating around in computer science research circles since around the year 2000. The META Group (now Gartner) published a research report in 2001 [1], where they described data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).

Every year since 1995, Gartner has produced an annual deliverable called the "hype cycle", which gives an idea of the technologies that survive the market hype and have a potential to become a part of daily life.

Formally, the Gartner Hype Cycle provides a graphic representation of the maturity and adoption of technologies and applications, and how they are potentially relevant to solving real business problems and exploiting new opportunities. This graph has five regions: *Innovation Trigger* (i.e. when potential technology breakthrough kicks off), *Peak of Inflated Expectations* (i.e. success stories through early publicity), *Trough of Disillusionment* (i.e. waning interest), *Slope of Enlightenment* (i.e. when 2nd & 3rd generation products appear) and *Plateau of Productivity* (i.e. when mainstream adoption starts).

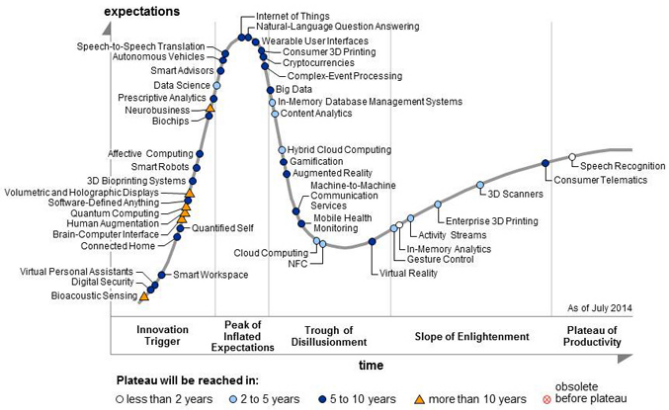

Fig. 2: Gartner Hype Cycle 2014.

In the 2014 Gartner Hype Cycle (Figure 2), Big Data is shown to be entering the Trough of Disillusionment, i.e. after a lot of discussion about the phenomenon and its promise, a general reduction in interest about Big Data overtakes the marketplace as experiments and implementations fail to deliver and or under-deliver. In this phase, investments continue only if the surviving players improve their products to the satisfaction of early adopters.

However, in the 2015 Gartner Hype Cycle (Figure 3), Big Data is completely gone. Given, the sustained high search activity (Figure 1), this may mean that the most talked about Big Data related technologies are now into practice and no more a hype.

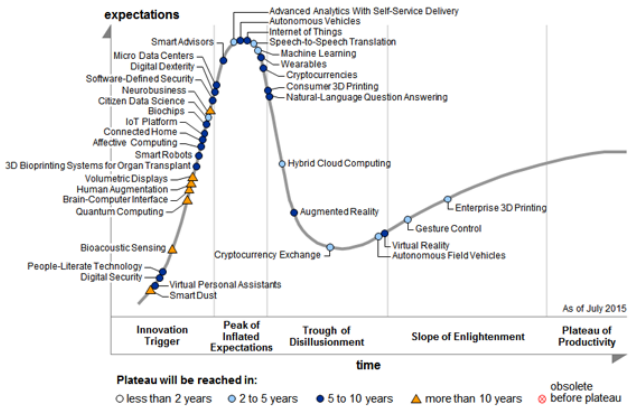

Fig. 3: Gartner Hype Cycle 2015.

This conclusion confirms that since the initial exploration of Big Data opportunities, the term and its associated technologies have been embraced, and are most likely embedded in the industry.

This adoption has progressed with little regard for the privacy implementations that must be included in Big Data technology stacks; as the benefits of Big Data far outweigh the potential commercial harm for businesses. This is the reason why the authors decided to explore this topic.

We first present the basics on Big Data (section 2). Then, in section 3, we outline the process that privacy professionals take when protecting data sets. In section 4, we model both the process of creating big data, i.e. information integration, and the process of protecting the privacy of the subjects specified in the Big Data sets; making observations about the model along the way. Then we discuss the impact of those observations in section 5. Finally, we present related work (section 6) and conclude (Section 7).

II. Big Data

Gartner, and most computing practitioners and researchers, still use the "3Vs" model as the basis for explaining the concept of Big Data [2]. In 2012, Gartner updated its definition, which now states that: "Big Data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [3].

Though, there is no consensus on a definition of the concept, the above characterization is the closest that the field has come to one and will be the foundation for this paper.

Over the last few years, the additional dimensions of value and veracity have been embraced, by many industry players, as being important additional characteristics of Big Data. So, currently, the "5Vs" are viewed as critical differentiators of Big Data versus regular data.

Formally, *volume* refers to the fact that organizations are collecting petabytes, exabytes, zettabytes and even yottabytes worth of data from multiple sources, e.g. sensors, social media, smart digital devices, etc. *Velocity* refers to the fact that these large volumes of data must be processed and analyzed as they stream into and out of an organization, i.e. in real-time, to extract value from it. *Variety* refers to the fact that this data is in various forms of organization, i.e. structured, semi-structured or unstructured, and in a number of modalities, i.e. free text, video, audio, sensor data, logs, etc. *Value* refers to the fact that this data is used to generate revenue through insight and influence spending, cost-saving strategies and optimizations. *Veracity* refers to the fact that the data (and data sources) must be trustworthy, as they impact critical choices.

We purport that in its simplest case, Big Data is the result of the convergence, i.e. integration, of large amounts of valuable, diverse, variegated data, where the schema of each constituent data set is a much smaller set than the data it describes (to the tune of several orders of magnitude), and where this data must be manipulated and analyzed very quickly.

It is collaboration that enables this convergence and makes new and exciting discoveries and innovations possible.

III. Current Privacy Preservation Practices

It is standard practice for privacy researchers and professionals who are tasked with analyzing data and transforming the sensitive portions of it (whether by anonymization, pseudonymization or other mechanisms) to create privacy-preserving versions of data; by partitioning the attributes of the data into three main categories: 1) sensitive data - can be used to identify an individual or data record, 2) quasi-identifiers - together with auxiliary information can be used to identify individuals and or records, and 3) benign data - is thought of as non-identifying or non-sensitive.

Practitioners tend to use government or industry guidelines in the data attribute classification process, e.g. the de-identification guidelines provided by the US Department of Health and Human Services (HHS) in the Health Insurance Portability and Accountability Act (HIPAA) [4] and the anonymization guidelines articulated in the code of practice document produced by the UK's Information Commissioner's Office (ICO), in line with the European Union (EU) Data Protection Directive [5]. Practitioners often do an initial analysis of the data, in the context of publicly available information, to determine the attributes that can be used to re-identify records, i.e. quasi-identifiers.

After this bucketing is performed, either the same or multiple transformative algorithms are applied to sensitive data and quasi-identifiers in the corpus to transform the data set into a safer, i.e. more privacy-preserving, form.

For example, let's examine a simple, non-normalized database table for contacting a services company, which we will refer to as *contact_us*.

The attributes of the *contact_us* table are:

1. *Name*: this is the name of the email's sender
2. *Email*: this is the email address of the sender
3. *City*: this is the city the sender lives in
4. *State*: this is the state the sender lives in
5. *Zip*: this is the sender's zipcode
6. *ID*: this is the unique identifier for the message
7. *Loc*: this is the folder location where the message and its attachments are stored
8. *Subject*: this is the subject of the message
9. *Message*: this is the message contents
10. *Date*: this is the date that the message was sent.

In the initial step, a privacy professional, leverages guidelines and an analysis of the data values to place attributes into one of the three buckets: 1) sensitive (*S*), 2) quasi-identifier (*Q*), and 3) the rest that are not sensitive and not quasi-identifiers (*R*).

For the *contact_us* table, a highly-probable and possible partitioning is:

S={*Name*, *Email*, *ID*}
Q={*City*, *State*, *Zip*, *Date*}
R={*Loc*, *Subject*, *Message*}

In the second phase, the practitioner determines the privacy-preservation algorithms to be used to transform the sets *S* and *Q*.

For this example, let's assume that sensitive data will be encrypted using the AES algorithm, using a key of arbitrary length – say 256 bits, and that quasi-identifiers will be hashed with the SHA-3 function.

After the functions have been applied, the privacy of the resulting data set is assumed to be higher than the privacy of its prior state. Stated another way, the privacy risk of the new state is lower than the privacy risk of the old state.

The underlying discussion on the correctness of assigning a particular attribute to a given category, the dynamic nature of an attribute over time and in the context of emerging data and thus, the fluidity of data and metadata are all topics that are out of the scope for this paper.

Also not included in the purview of this paper are the issues surrounding the correctness and utility of the algorithms employed to improve privacy and their appropriateness in being applied to various types of data and in particular contexts.

We take as a given that the privacy professionals correctly and consistently use guidelines to classify columns of data into various distinct classes and then appropriately apply algorithms to particular segments in order to protect sensitive items.

For our model – detailed in the next section - we assume the current state of affairs in the industry. We also choose the volume dimension of Big Data as our starting point; as it represents the simplest use case.

IV. The Model

Assume a data set, d_1, which is vertically segmented into a set of three types of data attributes or descriptors: s_1 – the set of sensitive attributes, q_1 – the set of quasi-identifiers, r_1 – the set of non-sensitive attributes that are not quasi-identifiers.

$$d_1 = \{s_1, q_1, r_1\}$$

We define T as a family of privacy transformation functions, where any member function can be applied to d_i to transform it into its privacy-preserving form, i.e.

$$\{t_1, \dots, t_n\} \in T \mid n \in \mathbb{N}$$

$$d_1^P = T(d_1) = \{t_a(s_1), t_b(q_1), r_1\} \mid t_a, t_b \in T$$

We assume that different transformations may be applied to both sensitive attributes and or quasi-identifiers. The information that is deemed innocuous is presumed to not require transformation.

We also define PR_i as a Privacy Risk function, a probability distribution function, which quantifies the risk of a data set, i.e.

$$PR_1(d_1^P) = \omega \mid \omega \in \mathbb{R}, 0 \leq \omega \leq 1$$

$$PR_1(d_q^P) = \omega \mid \omega \in \mathbb{R}, 0 \leq \omega \leq 1 \ and \ q \in \mathbb{N}$$

If PR_1 equals to 1, then it assumed that the data set is not private. The implication here is two-fold. Firstly, that individual record owners can be easily identified in the data set. Secondly, that individuals are completely exposed to a breach of some kind.

If PR_1 equals to 0, then it is assumed that the data set is completely private. This means that the identity of each person in the data set is completely protected and that even in the event of a breach incident the individual has little to be concerned about.

Assume that there is a universal set of privacy risk score functions, *PR*. For every privacy transformation function in this set, there is an associated privacy risk function.

$$\{PR_1, \dots, PR_n\} \in PR \mid n \in \mathbb{N}$$

$$((\forall t_a \in T) \rightarrow (\exists PR_a \in PR)) \mid a \in \mathbb{N}$$

An example of this co-existence of privacy transformation and risk function can be observed with the *k*-anonymity algorithm [6], i.e.

$$\exists t_a = (k - anonymity) \text{ such that } PR_a(d_b^P) = \frac{1}{k} \mid a, b, k \in \mathbb{N}$$

For typical information integration scenarios, there is at least a second data set, d_2, which is also vertically segmented and where the segments may or may not have common attributes as the segments in d_1.

$$d_2 = \{s_2, q_2, r_2\} \text{ such that}$$

$$((s^{\cap} = (s_2 \cap s_1))_{\wedge}(q^{\cap} = (q_2 \cap q_1))_{\wedge}(r^{\cap} = (r_2 \cap r_1))) \rightarrow ((|s^{\cap}| \geq 0)_{\wedge}(|q^{\cap}| \geq 0)_{\wedge}(|r^{\cap}| \geq 0))$$

For simplicity, we define *S* as the global set of all sensitive attributes, *Q* as the set of all quasi-identifiers and *R* as the set of all non-sensitive, non-quasi-identifiers.

$$S = \{s_1, \dots, s_n\}, Q = \{q_1, \dots, q_n\}, R = \{r_1, \dots, r_n\} \mid n \in \mathbb{N}$$

Where s_1 is the set of sensitive attributes from the first data set, s_n is the set of sensitive attributes from the *n*th data set, q_1 is the set of quasi-identifying attributes from the first data set, q_n is the set of quasi-identifying attributes from the *n*th data set, etc.

Please note that we assume that the cardinality of each set may vary. For our notation, this would mean that some of the descriptors may contain NULL values.

We define the data integration operator function, $\oplus$, which combines *n* data sets, based on a join criteria, *j*, which is a set of attributes common to all the data sets being integrated.

$$D = \oplus_j (d_1, \dots, d_n) = \{s_1, \dots, s_n, q_1, \dots, q_n, r_1, \dots, r_n\} \mid ((j \subseteq d_1)_\wedge \dots_\wedge (j \subseteq d_n))$$

The integrator function mimics the natural join operation in standard relational algebra. However, we make it an abstract function that can be used on well-described data, whether they are stored in relations or not.

The set *D* is the composite data set of all the constituent data sets that we integrated, i.e. *D* is a *Big Data* set. We refer to each of the data set parameters passed to the integrator function as operands or operand sets.

Given *n* privacy transformation functions, $t_1 \dots t_n$, where each transformation is different and applied to a corresponding data set, we assert that applying the integrator function on the privacy-preserving equivalents of the operand sets will mean that the join operators must be a subset of the non-sensitive non-quasi-identifiers.

$$D^P = \oplus_j (d_1^P, \dots, d_n^P) = \{t_1(s_1), \dots, t_n(s_n), t_c(q_1), \dots, t_z(q_n), r_1, \dots, r_n\} \mid ((\forall t_a, t_b \in T)_\wedge (t_a \neq t_b)) \rightarrow ((j \subseteq r_1)_\wedge \dots_\wedge (j \subseteq r_n))$$
$$a, b, c, n, z \in \mathbb{N}$$

If some or all of the transformation functions are the same, then the join set may also be a subset of the sensitive data and quasi-identifiers.

$$D^P = \oplus_j (d_1^P, \dots, d_n^P) = \{t_1(s_1), \dots, t_n(s_n), t_c(q_1), \dots, t_z(q_n), r_1, \dots, r_n\} \mid ((\forall t_a, t_b \in T)_\wedge (t_a = t_b)) ((j \subseteq s_1)_\wedge \dots_\wedge (j \subseteq s_n))_\vee ((j \subseteq q_1)_\wedge \dots_\wedge (j \subseteq q_n))_\vee ((j \subseteq r_1)_\wedge \dots_\wedge (j \subseteq r_n)) \quad a, b, c, n, z \in \mathbb{N}$$

The privacy risk of the *Big Data* set is the privacy risk of the combined constituent data sets.

$$PR(D^P) = PR\left(\oplus_j (d_1^P, \dots, d_n^P)\right) where\ n \in \mathbb{N}$$

Intuitively, the privacy risk of a Big Data set, which is made up of *n* smaller data sets, is the privacy risk of knowing an arbitrary data attribute(s) given all the prior attributes.

Lemma:

PR is monotonically increasing

i.e. $PR\left(\oplus_j (d_1^P, \dots, d_n^P)\right) \geq PR\left(\oplus_j (d_1^P, \dots, d_{n-1}^P)\right)$

For any arbitrary sequence of datasets, $d_1, \dots, d_n$, and $\oplus_j$ integrator operator function on join criteria $j, \exists k \subset j$ where k is a subset of the join criteria *j* that excludes $d_m \mid \forall m > n - 1\ where\ \ m, n \in \mathbb{N}$ dataset i.e. $((k \subseteq d_1)_\wedge \dots_\wedge (k \subseteq d_{n-1}))$.

$$\Rightarrow \oplus_j (d_1^P, \dots, d_{n-1}^P) \equiv \oplus_k (d_1^P, \dots, d_{n-1}^P)$$
$$\equiv \oplus_k (d_1^P, \dots, d_n^P) \subseteq \oplus_j (d_1^P, \dots, d_n^P)$$
$$PR\left(\oplus_j (d_1^P, \dots, d_{n-1}^P)\right) = PR\left(\oplus_k (d_1^P, \dots, d_{n-1}^P)\right)$$
$$= PR\left(\oplus_k (d_1^P, \dots, d_n^P)\right) \leq PR\left(\oplus_j (d_1^P, \dots, d_n^P)\right)$$
$$for\ any\ arbitrary\ Privacy\ Risk\ function, PR.$$

Thus, *PR* is monotonically increasing as $n \longrightarrow \infty$.

Theorem:

$$\lim_{n\to\infty} PR(D^P) = \lim_{n\to\infty} PR(\oplus_j (d_1^P, \dots, d_n^P)) \approx 1 \qquad (1)$$

Proof:

PR is a monotonic increasing function that is bounded above by 1. By monotone convergence, $\lim_{n\to\infty} PR(D^P) = sup(\Omega) \mid \forall \omega \in \Omega = 1$ since *PR* is a probability distribution function.

$$\lim_{n\to\infty} PR(D^P)$$
$$= \lim_{n\to\infty} PR\left(\oplus_j (d_1^P, \dots, d_n^P)\right)$$
$$= \lim_{n\to\infty} PR\left(d_n^P \mid (d_1^P, \dots, d_{n-1}^P)\right)$$
$$= \lim_{n\to\infty} \frac{PR\left(d_{n\wedge}^P (d_1^P, \dots, d_{n-1}^P)\right)}{PR(d_1^P, \dots, d_{n-1}^P)}$$
$$where\ n \in \mathbb{N}$$

As *n* tends to infinity, $PR\left(d_{n\wedge}^P (d_1^P, \dots, d_{n-1}^P)\right)$ tends to 1 and so does $PR(d_n^P)$. Thus,

$$\lim_{n\to\infty} \frac{PR\left(d_{n\wedge}^P (d_1^P, \dots, d_{n-1}^P)\right)}{PR(d_n^P)} \approx 1$$

In layman's terms, as you integrate more data sets into an existing data collection, the privacy risk score of the cumulative data set (which is currently assumed by most people to be privacy-preserving) increases to the point where there is no privacy.

Put even simpler, *Big Data* sets have a very high probability of not being private.

V. Discussion

This paper is a first attempt at shining the spotlight on a fundamental risk inherent in Internet collaboration, i.e. the privacy risks associated with the data sets used and produced during the interaction.

As Big Data is the typical starting point for online collaboration, we focused on Big Data privacy to highlight the validity of a basic assumption – that privacy is possible for Big Data sets.

In this discussion section, we will speak to possible future explorations based on the current model and then to the general insight from our formulation.

A. Model Implications

Assuming that the model is an accurate enough approximation of current privacy practices used and the Big Data formation process, there are a few natural considerations.

If one were to dive into the model, particularly equation (1), there is an interesting question that arises – "Is there a value for n at which the privacy risk becomes significant?" Simply put, how many data sets does it take to shift the entire collection from a safe to an unsafe state? This question assumes that such a threshold value can even be determined. Such a threshold may probably even be contextual and will vary based on the context of the data collection.

The convergence of PR to 1 as new data sets are added implies that the set of attributes used to join the new and prior data sets moves from benign & quasi-identifying to being "effectively identifying". This validates the dynamic nature of data attributes, but calls into question the notion that it is possible to do an initial classification of attributes into the initial three buckets that will stand the test of time, or rather integration.

Even though it is the authors' viewpoint is that any piece of data may, at a point in time, be situationally or contextually sensitive, we did not question the validity of the nature of data in this current paper. Instead, we worked within the confines of current best practice. Determining the data attributes that privacy professionals place into each category is definitely further of further examination.

The model focuses sharply on the volume component of a Big Data set, i.e. the size of the data set, n. The authors viewed volume as the base case for Big Data collections. However, we foresee other models that factor in the rate of integration of newer data sets (velocity), the modality of the data sets (variety), the worth of the data set (value) and the trustworthiness of the data set and source (veracity).

Let's elevate our perspective and examine the macro-level observations from this exposition.

B. General Observations

In our model, we include the notion that there is some attempt to protect privacy in the data sets. We assumed that participants who are hosting, combining and collaborating on data sets are good stewards and try to take the necessary steps to reduce privacy risk.

From this "best case" model, it is seen that safeguarding privacy when it pertains to Big Data is close to impossible, probabilistically.

Imagine the cases where some of the operand data sets are not protected and are merged with data that is thought to be protected. In these scenarios, the privacy risk most likely increases at a faster rate towards an unsafe state.

The model and the proof are purposefully simple. The intention is to provide a basic formulation to demonstrate a simple and intuitive result. The authors want this conclusion to be the starting point for a robust discussion on real solutions that enable Big Data privacy – ones that view all data as sensitive, provide cryptographically hard ways to perform privacy transforms and that enable the processing of that data in its secure state.

VI. Related Work

This work builds on the seminal work at Gartner [1], where the concept of Big Data was defined and its dimensions articulated and explored.

The policy and practice elements of this paper builds on HIPAA [4], and the EU Data Protection Directive [5].

The algorithms underlying k-anonymity [6], Hippocratic database technology [7-10] and differential privacy [11] provide the technical underpinnings of the model we presented.

The current set of privacy models, e.g. Fischer-Hübner and Ott [12], Gilburd et. al. [13], Hermans et. al. [14], and Bohli et. al. [15], focus on describing the classical elements of privacy policies, i.e. purpose, disclosure, use, consent, etc, and on using access control frameworks to implement these elements in service of privacy protection. In contrast, this paper assumes that these specification and implementation models exists and work as defined. We focus on working at a higher level of abstraction; examining the effect on risk of combining these "private" data sets.

Popa et. al. [16] propose a mechanism for performing encrypted queries on encrypted Big Data repositories. This is one of the approaches that naturally supports the next steps for this paper.

VII. Conclusion

As researchers in a field that is fundamentally changing the collective way of life for citizens across the globe, our community has the responsibility to ensure that the technologies that we create and provide is being built with care and used ethically on everyone's behalf and for everyone's benefit.

Our current focus on collaboration masks a fundamental flaw at the core of our infrastructure – the privacy of the Big Data sets upon which these interactions are based.

In the field today, there is general acceptance of the notion that Big Data privacy is possible through the application of simple privacy-preserving transformation algorithms. This assumption may not only be deceptive to the general public, but also holds the potential to unravel an entire industry and invalidate a field of study, if we do not have the will to critically analyze our systems, processes, and self-interest.

In this paper, we focus on 1) describing the "as-is" state of the data privacy protection practice, and on 2) modelling the core of what is Big Data. We then layered the two worlds together using a probabilistic framework and took that framework to its obvious, natural conclusion. The end result – given the current model, there is no privacy when it comes to Big Data.

This is the start of a critical discussion and introspection – one that we hope the community will engage in.

ACKNOWLEDGMENT

Special thanks to Jeffrey Chen and Sean Thorpe that provided feedback on earlier versions of this paper.

REFERENCES

[1] Laney, D. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. February 6th, 2001. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf on May 1st, 2013.

[2] Beyer, M. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. June 27, 2011. Retrieved from https://www.gartner.com/newsroom/id/1731916 on May 1st, 2013.

[3] Laney, D. "The Importance of 'Big Data': A Definition". Gartner. June 21, 2012. Retrieved from https://www.gartner.com/id=2057415 on May 1st, 2013.

[4] US Department of Health and Human Services. "Health insurance portability and accountability act". Retrieved from www.hhs.gov/hipaa/ on October 7th, 2016.

[5] European Commision. "Protection of personal data". Retrieved from ec.europa.eu/justice/data-protection/ on October 7th, 2016.

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

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

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

[9] Agrawal, R., et al. "Auditing compliance with a hippocratic database." *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004.

[10] Agrawal, R., Bird, P., Grandison, T., Kiernan, J., Logan, S., & Rjaibi, W. (2005, April). Extending relational database systems to automatically enforce privacy policies. In *ICDE* (Vol. 5, pp. 1013-1022).

[11] Dwork, C. (2008, April). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation* (pp. 1-19). Springer Berlin Heidelberg.

[12] Fischer-Hübner, S., & Ott, A. (1998, October). From a formal privacy model to its implementation. In *Proceedings of the 21st National Information Systems Security Conference* (pp. 5-8).

[13] Gilburd, B., Schuster, A., & Wolff, R. (2004, August). k-TTP: a new privacy model for large-scale distributed environments. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 563-568). ACM.

[14] Hermans, J., Pashalidis, A., Vercauteren, F., & Preneel, B. (2011, September). A new RFID privacy model. In *European Symposium on Research in Computer Security* (pp. 568-587). Springer Berlin Heidelberg.

[15] Bohli, J. M., Sorge, C., & Ugus, O. (2010, May). A privacy model for smart metering. In *2010 IEEE International Conference on Communications Workshops* (pp. 1-5). IEEE.

[16] Popa, R. A., Redfield, C., Zeldovich, N., & Balakrishnan, H. (2011, October). CryptDB: protecting confidentiality with encrypted query processing. In Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles (pp. 85-100). ACM.