View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by AIS Electronic Library (AISeL)

Privacy-preserving Distributed Analytics: Addressing the Privacy-Utility Tradeoff Using Homomorphic Encryption for Peer-to-Peer Analytics

Research-in-Progress

Prasanta Bhattacharya
Department of Information Systems
National University of Singapore
Singapore 117417
prasanta@comp.nus.edu.sg

Tuan Q. Phan
Department of Information Systems
National University of Singapore
Singapore 117417
phantq@comp.nus.edu.sg

Linlin Liu
Department of Information Systems
National University of Singapore
Singapore 117417
a0078051@u.nus.edu.sg

Abstract

Data is becoming increasingly valuable, but concerns over its security and privacy have limited its utility in analytics. Researchers and practitioners are constantly facing a privacy-utility tradeoff where addressing the former is often at the cost of the data utility and accuracy. In this paper, we draw upon mathematical properties of partially homomorphic encryption, a form of asymmetric key encryption scheme, to transform raw data from multiple sources into secure, yet structure-preserving encrypted data for use in statistical models, without loss of accuracy. We contribute to the literature by: i) proposing a method for secure and privacy-preserving analytics and illustrating its utility by implementing a secure and privacy-preserving version of Maximum Likelihood Estimator, "s-MLE", and ii) developing a web-based framework for privacy-preserving peer-to-peer analytics with distributed datasets. Our study has widespread applications in sundry industries including healthcare, finance, e-commerce etc., and has multi-faceted implications for academics, businesses, and governments.

Keywords: Big Data Analytics, Privacy-preserving Analytics, Homomorphic Encryption, Distributed Architecture.

Introduction

We live in an information age and performing analytics on data is becoming increasingly valuable to businesses (Chen et al. 2012). However, data may be fragmented and held by multiple organizations - limiting its usefulness. Privacy, security, and legal issues, in addition, may restrict data sharing (Goldfarb and Tucker 2012; Hann et al. 2007). Responding to the recent concerns over data privacy, Governments too are introducing stricter policies and laws (Information Shield 2011). Researchers in the field of Information Systems as well as related businesses are thus faced with the tradeoff between deriving useful insights from data while respecting individual rights to information privacy. In our study, we attempt to address this tradeoff by designing and developing a secure distributed analytics system which parties can use to effectively share and analyze datasets without compromising on either accuracy or privacy.

In this research in progress paper, we propose and build a system to perform peer-to-peer secure analytics using homomorphic transformations on data from two or more parties. While recent research on homomorphic encryptions have hinted at the potential for this method to be used in cloud computing environments (Hall et al. 2011), ours is the first to design and implement a workable peer-to-peer (P2P) solution that leverages this concept. We specify an efficient communication protocol using REST, for performing complex mathematical operations, including multiplicative inverse and statistical operations, including linear and logistic regressions and numerical optimizations in a computationally acceptable manner. Our system has wide-spread benefits in many industries where data secrecy and customer privacy are major concerns, like the finance, healthcare, retail, and telecommunications industries. Our proposed method will also enable researchers and academics in marketing, information systems and related disciplines to investigate questions pertaining to information sharing, data security and privacy-preserving analytics from a fresh perspective (Xu et al. 2012; Bélanger and Crossler 2011; Sheth 2011).

Our system provides a number of advantages and capabilities over existing solutions which use data anonymization or privacy-preserving tools. We argue that previous systems are often inadequate, risky and prone to a single-point-of-failure, or popular attack. While most public datasets revealed by companies are anonymized to protect user privacy, researchers hint that perfect anonymization is not possible without damaging the utility of the data (Sweeney 2010; Narayanan and Shmatikov 2008). In fact, there is increasing evidence, both anecdotal and grounded, that it's fairly easy to de-anonymize popular datasets by comparing data points across different datasets (Ohm 2010; Narayanan and Shmatikov 2008). Others argue in favor of the efficacy of existing anonymization approaches (Cavoukian and Emam 2011). We contend, however, that in light of such conflicting information on anonymization strategies, organizations would be increasingly reluctant to use such approaches. The efficacy of privacy-preserving tools is questionable too as most of them rely on "trusted" third parties. Sharemind is a good example of this (Bogdanov et al. 2008). A clear shortcoming of this system is the constant risk of data disclosure if the data miners are corrupt or are compromised.

We address these shortcomings of existing secure analytics solutions by proposing a system that leverages the well-established cryptographic approaches. The Paillier's encryption scheme that we describe in this paper draws upon homomorphic properties to allow arbitrary computations on fully encrypted data. Thus, our P2P system provides stronger privacy guarantees than simple anonymization techniques while not relying on any single third party for performing the secure computations. As with most homomorphic encryptions, the elephant in the room is the efficiency of computations. However, in our project,we introduce certain design improvisations in our security protocols to make them computationally acceptable solutions for use in industry and academia.

In the next section, we briefly look at some previous work on information privacy and privacy-preserving data analysis. Following this, we introduce the two-party protocols which we have designed for various mathematical and statistical operations. Next, we present some preliminary evaluation results and conclude with a brief discussion on the research plan ahead.

Background

Individual's Information Privacy Debate

Information privacy has been an active area of interest among Information Systems researchers and practitioners. We draw heavily from two meta-analyses by Bélanger and Crossler (2011) and Smith et al. (2011) to discuss and guide our review of the current discourse in this popular domain. In their analysis, Smith et al. note that most of the previous research in information privacy spanning Economics, Marketing, Law, Philosophy and Information Systems disciplines have attempted to answer one of the following three questions about privacy: (i) *What is (and is not) privacy and how is it different from the notion of security?* (ii) *What is the relationship between privacy and other related constructs?* (iii) *To what extent does context matter in the relationship between privacy and other constructs?* (i.e. how generalizable are privacy related findings across industries and environments?). Several studies have attempted to discuss the first from philosophical, psychological, sociological and legal perspectives, with limited consensus (Solove 2006; Westin 1968). This has led to a stark increase in several competing theoretical frameworks, with often conflicting empirical evidence (Bélanger and Crossler 2011; Siponen 2005). At the heart of these discussions on information privacy lies an ongoing debate between the idea of privacy as a general right (Warren and Brandeis 1890; Rosen 2012; Bennet 2012) and as a commodity (Campbell and Carlson 2002; Davies 1997; Laudon 1996).

In addition, a related stream of literature illustrates the idea of a *privacy calculus* by assuming that individuals face a tradeoff between the costs and benefits of privacy disclosure, and that this tradeoff is salient in guiding the user's behavior in privacy decisions (Klopfer and Rubenstein 1977; Laufer and Wolfe 1977; Posner 1981; Stone and Stone 1990; Chellappa and Sin 2005; Hui et al. 2006; Xu et al. 2009). In other words, an individual's decision to reveal personal information depends on the outcome of a rational cost-benefit analysis of disclosing this information (Dinev et al. 2006; Krasnova and Veltri 2010). More recent studies have pointed out that while higher privacy is clearly desired by end-users (Goldfarb and Tucker 2012), it might reduce the quality of services provided to them e.g. poor targeting of online ads, and thus adversely affect their preferences towards the service (Goldfarb and Tucker 2011). However, this reduction might be countered by an increase in the willingness of the users to use the service, due to the added privacy guarantees (Tucker 2014). Evidently, researchers have proposed certain information-theoretic frameworks to better quantify these risks and benefits of data disclosure (Sankar et al. 2013; Brickell and Shmatikov 2008; Li and Li 2009; Rastogi et al. 2007).

Privacy-preserving Analytics

Existing approaches to preserving security and privacy of data involve use of asymmetric key encryption algorithms and the partitioning of cloud storage into unsecure or "unsafe" zones, which are essentially databases containing unencrypted data and "safe" zones, which are storage locations deemed to be secure to external breaches (Yu and Liu 2007). Most data analysis happens by first transporting encrypted data from the unsafe to the safe zones and then decrypting it before the analysis starts. However, the efficacy of this approach depends on the safety of data in "safe" zones. Recent reports suggest that these might also be vulnerable to external threats (Wilson and Ateniese 2014; Purewal 2014).

Another approach originating from the statistical sciences looks into understanding how best to perform analysis without compromising security and privacy (Duncan et al. 2011), through data masking (Duncan and Stokes 2009) or data-noising strategies. This approach reduces the problem to that of extracting usable information from noisy data (Chen et al. 2009; Duncan and Stokes 2009). While this approach is fairly robust to standard security attacks like the man-in-middle and SQL injection, the accuracy of the analysis result often suffers as a function of the amount of noise introduced in the initial data (Agrawal and Srikant 2000). Moreover, repeated observations of the encrypted data might also reveal insights about the usage pattern and result in possible de-anonymization of the data. In rare cases, such usage patterns can be traced back to unique individuals. A number of more recent studies have also looked at secure single- and multi-party computation techniques, similar to the one described in this study, but lack any performance benchmarks on real-world data (Du et al. 2004; Chaudhuri and Monteleoni 2008; Zhang et al. 2012)

In the current study, we address the limitations of the above approaches by using a class of homomorphic encryption algorithms, specifically the Paillier's Cryptosystem (Paillier 1999). As introduced earlier, homomorphic encryption is a form of asymmetric-key data encryption which permits certain types of computations to be carried out on encrypted data to generate an encrypted result which, when decrypted, gives the same result as the one obtained by performing the computation on unencrypted data (Rivest et al. 1978; Gentry 2009). The intuition behind this is that structure-preserving transformations (i.e. encryption) of the data would behave similar to the actual data, with certain permitted mathematical operations (e.g. additions, multiplications etc.). This avoids the need to decrypt data before performing any data analysis.

Secure Two-Party Protocols

Our method builds upon Hall et al.'s (2011) initial proposed framework which used Paillier's homomorphic encryption (Paillier 1999). The concept behind these protocols hinges on the basic idea of a private data "share." Consider the model specification for a linear regression model:

$$y = X\beta + \varepsilon \tag{1}$$

In the model above, X represents design matrix and y represents response matrix. Since we are performing analysis on data from multiple parties, each party would provide several rows or columns of data. The complete dataset, however, would not be known to any party at any point in time. Assume there are n parties and the design matrix of party i is represented by X_i and the response matrix by y_i, then $X = X_1 + X_2 + \cdots + X_n$ and $Y = Y_1 + Y_2 + \cdots + Y_n$. Here we call X_i and y_i shares from each party. If the sum of two numbers$a$ and b is c, we can also call a and b shares of c. We extend this idea of distributed shares to a secret share scheme as illustrated later in this paper, wherein the intermediate results for any statistical analysis would not be stored with any one party. Instead, these results would be split into secret shares and stored with multiple parties, thereby protecting the system from any data leakage. Next, we describe the basic properties of Paillier's encryption scheme which we leverage in our system.

Paillier's Cryptosystem

We summarize the homomorphic properties of the Paillier's cryptosystem below (Paillier 1999; Hall et al. 2011). Variables a, b and c represent integers from the set $\mathrm{Z_n} = \{0, 1, \ldots, \mathrm{n}\}$, and $\mathrm{E_n(a)}$ represents the encrypted value of a, where n is the public key, then the cryptosystem has the following properties:

$$E_n(a) \cdot E_n(b)\ mod\ n^2 = E_n(a + b) \tag{2}$$

$$E_n(a)^c\ mod\ n^2 = E_n(a \cdot c) \tag{3}$$

Since the above cryptosystem can only encrypt integers in the set $Z_n = \{0, 1, \ldots, n\}$, we map the integers to real numbers for application in our study.

$$f: \mathbb{Z}_n \to \mathbb{R}, f(a) = M^{-1} \begin{cases} a & a \leq \frac{n}{2} \\ a - n & a > \frac{n}{2} \end{cases} \tag{4}$$

In this method, n is the public key and M is a constant which determines the precision of the fractional representation. Further, $-\frac{n}{2M}$ and $\frac{n}{2M}$ are the smallest and largest number represented using this method.

Two-Party Protocols for Mathematical Operations

To compute the secure protocols for linear and logistic regression models we first need secure sub-protocols for the intermediate mathematical operations including addition, multiplication and multiplicative inverse. Past studies have implemented secure protocols for matrix addition, multiplication and inverse (Karr et al. 2005; Hall et al. 2011).However, since we intend to implement protocols for logistic models, we extend the design to support new sub-protocols for operations including sum-of-products and exponentiation. We also note that the previous techniques to compute multiplicative inverse using Schur-Newton approaches (Guo and Higham 2006) are practically inefficient with large data. As a result, we design and implement a new protocol for performing multiplicative inverse, which is computationally more efficient, given the networking cost. In the following subsection, we highlight the

protocols for these new contributions viz. secure protocol for sum of products, secure protocol for exponentiation and an improved secure protocol for multiplicative inverse.

Secure Two-Party Protocol for Computing Sum of Products

Past research have provided protocols for computing simple products of integers, where each party knows one of two numbers, $x_1$and x_2, and the desired result is to compute $x_1 * x_2$. We extend this simple system to also perform sum of products. For instance, assume $X = x_1 + x_2$ and $M = m_1 + m_2$, where x_1 and m_1 are shares from party one and x_2 and m_2 are shares from party two. If we need to compute the value of $X * M$ as is the requirement with some non-linear models we propose a protocol to do so. It is important to emphasize at this point that the outputs from these mathematical protocols are essentially intermediate results for the statistical estimations we illustrate later. Thus, these intermediate results are not stored "as is" with any one party, but instead, stored as secret data shares which when combined produce the intermediate result. These data shares, denoted by k_1 and k_2 in Fig 1 below, have no informative property by themselves and thus offer no opportunity for data leakage.

Input: Party one has the private key to an instance of Paillier's encryption scheme, and shares x_1 and m_1. Party two knows the corresponding public key n, and shares $x_2$and m_2. $X = x_1 + x_2$, $M = m_1 + m_2$.
Step 1 Party one computes x_1m_1 locally and sends the encrypted values of x_1m_1, m_1 and x_1 to party two.
Step 2 Party two computes x_2m_2 locally. Then, party two applies the homomorphic property (3) of Paillier's cryptosystem to compute the encrypted value of $k = x_1m_1 + x_1m_2 + x_2m_1 + x_2m_2$. Thus, k is the product of X_s and M_s.
Step 3 Party one and party two apply the standard two-party protocol for computing simple products (Hall et al. 2011) to compute the shares of k, that is k_1 and k_2.
Output: Party one outputs share k_1 and party two outputs share $k_2$such that $k1 + k2 = x_1m_1 + x_1m_2 + x_2m_1 + x_2m_2$

Figure 1. ***Secure Two-Party Protocol for Computing Sum of Products***

Secure Two-Party Protocol for Computing Exponentiation

In this section, we use the homomorphic properties of the Paillier's encryption system to compute exponential functions when the shares are known. This will be a necessary operation for logistic regression as we demonstrate later in the paper. This protocol is based on the property of the exponentiation function, $n^{a+b} = n^a n^b$, so each party can compute n^a or n^b locally and then apply the standard two-party protocol (Hall et al. 2011) for computing products to generate the shares of n^{a+b}. For example, if we want to raise a number m to the power of $X = x_1 + x_2 \; mod \; n$, when m is known to both parties and x_1 and x_2 are shares of the exponent, we can use the protocol in Fig. 2 below to compute the shares of $m^{x_1+x_2}$.

Input: Party one has the private key to an instance of Paillier's encryption scheme, and share x_1, party two knows the corresponding public key n, and share x_2. Also, $X = x_1 + x_2 \; mod \; n$.
Step 1 Party one uses the real number mapping method to compute the real numbers that $x_1$and m are mapped to, which are represented by $f(x_1)$ and $f(m)$ respectively. Then, party one computes $f(m)^{f(x_1)}$, and uses the real number mapping method to map its result to a_1. Then, the encrypted value of a_1 is sent to party two.
Step 2 Party two computes $f(m)^{f(x_2)}$ and maps the result to a_2. Party two applies the homomorphic property (3) of Paillier's cryptosystem to compute $En(a_1a_2)$.
Step 3 The two parties now apply the standard two-party protocol for computing simple products to generate shares of the product of $a_1$and a_2.
Output: Party one gets output k_1 and party two gets output k_2 such that $k_1 + k_2 = m^{x_1+x_2}$.

Figure 2. ***Secure Two-Party Protocol for Computing Exponentiation***

Improved Two-Party Protocol for Computing Multiplicative Inverse

Inverting a number is a particularly important mathematical operation since it forms a key part of most statistical estimator functions, including linear and logistic regressions. Hall et al. (2011) suggest using the Schur-Newton method (Guo and Higham 2006) to compute the inverse of a number. However, this method is iteration-intensive and hinges on strong restrictions on the stopping criteria to achieve convergence. While this is acceptable for Hall et al. who use experimental simulations to demonstrate their protocols, it entails significant I/O delays for us since our distributed system is deployed on physical servers to emulate a real world P2P scenario. To resolve this concern, we develop a more I/O efficient protocol for inverting a number or a matrix. Assume that party one has the private key and party two has the public key, and that party one knows a number x_1 and party two knows a number x_2 such that $X = x_1 + x_2$. To compute $\frac{1}{X}$, the parties use the protocol in Fig. 3 to compute shares of the inverse of a number.

Input: Party one has the private key and party two has the public key, and that party one knows a number x_1 and party two knows a number x_2 such that $X = x_1 + x_2$.
Step 1 Party one sends the encrypted value of x_1 to party two.
Step 2 Party two applies the homomorphic property (2) of Paillier's cryptosystem to compute the encrypted value of $X = x_1 + x_2$. Then party two generates a random number r and applies the homomorphic property of Paillier's cryptosystem to compute the encrypted value of XrM, because the fixed point number mapping method will result in a stray factor M.
Step 3 Then, party two may take the encrypted value of XrM as an input of the protocol shown in Fig 1. After applying the two-party protocol for computing products, party one and party two would get shares k_1 and k_2 respectively, and $k_1 + k_2 = Xr$. Party two can now send k_2 to party one.
Step 4 Party one computes the sum of k_1 and k_2, which is the value of Xr.Then party one can apply the real number mapping method to get the real value of Xr, compute its multiplicative inverse, and then map the result to a integer representation of $\frac{1}{Xr}$.
Step 5 Party one and party two now apply the standard two-party protocol for computing simple products to compute the shares of the product of $\frac{1}{Xr}$ and r.
Output: Parties one and two get a fresh pair of data shares k_1 and k_2 such that, $k_1 + k_2 = \frac{1}{Xr} \times r = \frac{1}{X}$.

Figure 3. ***Improved Two-Party Protocol for Computing Multiplicative Inverse***

Secure Two-Party Protocols for Statistical Models

Using our extended framework, we implement secure-MLE (s-MLE), a secure protocol for logistic regression using maximum likelihood estimator and numerical optimization procedures. The binomial logistic model specification is as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{K} x_{ik}\beta_k \; i = 1,2,\dots,N \qquad (5)$$

In this model, we assume there are N records and K independent variables in the dataset. X is the design matrix, which is composed of N rows and K+1 columns. The elements in the first column of X are all 1. x_{ik} is an element of X on row i and in column k. β is a column vector with K+1 rows that represents the coefficient vector. Column vector π also has N rows, with each element π_i representing the probability of the dependent variable taking a value 1, given the observation on the i^{th} row. Our goal here is to estimate β in the above specification (5).The parameters in βcan be estimated with maximum likelihood estimation using the Newton Raphson (NR) optimizer, which helps to find the parameters for which the probability of the observed data is the greatest (Czepiel 2002). Although there are better methods than the Newton Raphson approach that would considerably lower the I/O load on our system, but for illustrative purposes, we retain the NR approach for the current paper. The matrix formulation of the NR iterative

function is as shown below where X and y are the design and response matrices respectively. W is a square matrix of order N, with $\pi_i(1-\pi_i)$ on the diagonal and zero elsewhere.

$$\beta^{s+1} = \beta^s + [X^T W X]^{-1} X^T (y - \pi) \quad (6)$$

Similarly, π_i can also be expressed in the format of matrix operations as follows:

$$\pi_i = \left(\frac{e^{J_i}}{1+e^{J_i}}\right) \quad (7)$$

Using our sub-protocols, the secure protocol for logistic regression can be designed as shown in Figure 4.

Step 1: We apply the extended two-party protocol for computing products (Fig. 1) to compute shares of Xβ. This gives us the shares of J_i, which is the i^{th}element of column vector Xβ. Next, we use the exponentiation protocol (Fig. 2) to compute shares of e^{J_i}, and the improved multiplicative inverse protocol (Fig. 3) to compute shares of $\frac{1}{1+e^{J_i}}$. Now, with shares of e^{J_i} and shares of $\frac{1}{1+e^{J_i}}$ as inputs to the extended two-party protocol for computing products (Fig. 1), each party gets a share of $\pi_i = \left(\frac{e^{J_i}}{1+e^{J_i}}\right)$. Next, we construct matrix shares of π since π_i are elements of this column vectorπ.
Step 2: Construct shares of W with $\pi_i(1-\pi_i)$ on the diagonal and 0 elsewhere. Since, $\pi_i(1-\pi_i) = \pi_i - \pi_i\pi_i$, the two parties apply the extended two-party protocol for computing products (Fig. 1) to compute shares of $\pi_i\pi_i$. Next, they utilize the homomorphic property (2) of Paillier's cryptosystem to compute the shares of $\pi_i - \pi_i\pi_i$.
Step 3: Each party applies the protocol from Fig. 1 to compute shares of $X^T W$, since shares of X and W are known to each party. In the same way, shares of $X^T W X$ are computed.
Step 4: The two parties now apply the protocol for matrix inversion (Fig. 3) to compute shares of $[X^T W X]^{-1}$.
Step 5: The two parties now use Paillier's homomorphic property to compute shares of $y - \pi$. Next, the shares of $y - \pi$ and shares of X^T are used as inputs to the protocol for computing products (Fig. 1) to compute shares of $X^T(y - \pi)$.
Step 6: Shares of $[X^T W X]^{-1}$ and $X^T(y - \pi)$ are now used as inputs to the protocol for computing products (Fig. 1) to compute shares of $[X^T W X]^{-1} X^T (y - \pi)$.
Step 7: For the first round of the NR iteration, shares of β^s are initialized to a zero vector. Each party's shares of β^s and $[X^T W X]^{-1} X^T (y - \pi)$ are added up locally to get a share of β^{s+1}. Next, the two parties check the difference between β^{s+1} and β^s for convergence. If not, they check whether β is tending to infinity, in which case, they initialize β with a new value and start from Step 1. Otherwise, both parties preserve the value of β and continue from Step 1.

Figure 4. ***Secure Two-Party Protocol for Estimating a Logistic Regression Model using s-MLE***

Runtime Evaluation

We are currently evaluating the suitability of our design artifacts (Hevner and Chatterjee 2010; Hevner et al. 2004) by analysing the runtime performance and scalability of our system to understand the protocol design limits and the effect of encryption delays. We present here some preliminary benchmark data from our evaluation tests. Table1 shows the runtime complexities of each protocol implemented in our system when performing analytics on a matrix X of i rows and j columns and matrix y of i rows and 1 column. Figure 5 shows the performance of our system with increasing length of the public key, which is a major contributor to encryption delays. The length of public key determines the "hardness" of our encryption scheme as well the range of numbers that our real number mapping method could represent. In our real number mapping method, M in formula (4) determines the precision of our number representation. In this implementation, $M = 2^{64}$ is used, so the smallest positive number we could represent is 2^{-64}, which is precise enough in most cases. According to (4), $-\frac{n}{2M}$ and $\frac{n}{2M}$ are the smallest and largest number that could be represented. Thus, using a 512-bit public key, the range of numbers we could potentially use is approximately between $-2^{446}$and 2^{446}. Therefore, unless the user wants to perform analytics on extremely large numbers or the user has an extremely high requirement on the precision of the analytics result, a 512-bit public key should be acceptable.

Table 1. Runtime Complexity	
Protocol	**Big-O**
Encryption/Decryption	$O(ij)$
$X^T X$	$O(ij^2)$
$X^T y$	$O(ij)$
$(X^T X)^{-1}$	$O(j^3)$
$X(X^T X)^{-1}$	$O(ij^2)$
$H = X(X^T X)^{-1} X^T$	$O(i^2 j)$
$e = (I - H)y$	$O(i^2)$

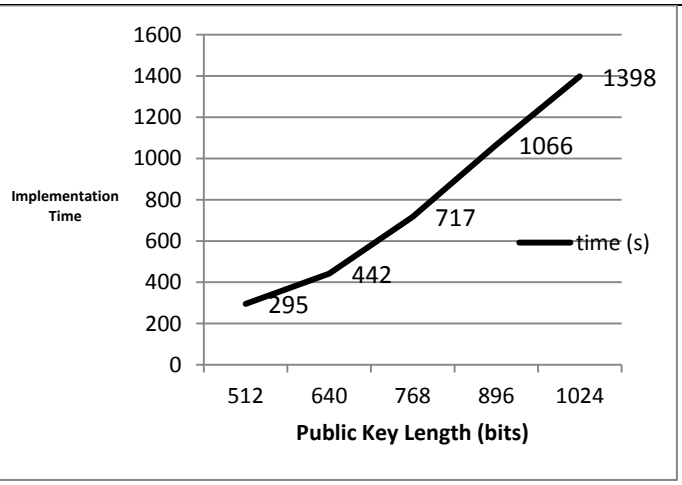

Figure 5. ***Regression delays (sec) with varying public key sizes***[1]

Conclusion and Future Work

In this paper we illustrate our initial efforts at proposing a new system for performing analytics on sensitive data that is held by separate parties e.g. separate organizations or separate research groups. While the benefits of performing analytics on combined datasets are well understood in academia and industry, there is no satisfactory method that allows us to do so in a secure and privacy-preserving fashion. As a result, researchers and practitioners in IS and related disciplines constantly balance a privacy-utility tradeoff when performing analytics on user data. We argue that contemporary techniques to deal with this problem compromise on either accuracy or privacy. While there have been some early inroads into using homomorphic techniques to solve this problem (Hall et al. 2011), considerable effort is still required to handle the communication and practical bottlenecks. The utility of existing approaches therefore remains limited due to the lack of a workable system design, absence of real world benchmarking data and a lack of focus on non-linear models that are most popular in information systems, computer science and statistics. Our study aims to address these gaps. The current paper highlights that our early attempts at doing so have been fruitful. While our system provides significant improvements over existing techniques which use anonymization and third-party privacy tools, we also extend past work on homomorphic encryptions by designing newer protocols and developing a fully-operational and web-based analytics system. Moreover, our system makes no apriori assumptions on the trustworthiness of the parties, their locations or the specific nature of the datasets.

As next steps, we are looking to extend our research on two fronts. First, we are running performance evaluation tests to identify computationally costly points with our existing secure protocols. For instance, knowing the delay-intensive operations would help us to not just make efficient protocols but to also make pointed recommendations to other security researchers on how to design computationally efficient protocols for big data. For this purpose, we are also developing parallelized versions of the secure protocols mentioned in this study that could be deployed on Graphics Processing Units (GPUs) for fast execution. On completion of this parallelization task, we would be able to provide both time-efficient as well as secure protocols for statistical computations. Second, we plan to continue designing protocols for diverse statistical models that provide increasing benefits to researchers and practitioners alike. In the present paper, for instance, we have showcased our design and implementation for logistic regression, a popular statistical modeling technique. Further work could look into other useful models like panel data models, duration models e.g. proportional hazard models, and clustering algorithms. This can potentially benefit both academic researchers as well as analytics professionals.

[1]Server configuration for all experiment devices: Processor: i7-2630QM 2GHz, RAM: 8 GB

References

Agrawal, R., and Srikant, R. 2000. "Privacy-preserving data mining," *ACM SIGMOD Record* (29:2)ACM, pp. 439–450.

Bélanger, F., and Crossler, R. E. 2011. "Privacy in the digital age: a review of information privacy research in information systems," *MIS Quarterly* (35:4)Society for Information Management and The Management Information Systems Research Center, pp. 1017–1042.

Bennett, S. C. 2012. "Right to Be Forgotten: Reconciling EU and US Perspectives, The," Berkeley Journal of International Law (30) (available at http://heinonline.org/HOL/Page?handle=hein.journals/berkjintlw30&id=163&div=&collection=).

Bogdanov, D., Laur, S., and Willemson, J. 2008."Sharemind: A framework for fast privacy-preserving computations," in *Computer Security-ESORICS 2008*, Springer, pp. 192–206.

Brickell, J., and Shmatikov, V. 2008. "The cost of privacy," in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08, New York, New York, USA: ACM Press, August 24, p. 70 (doi: 10.1145/1401890.1401904).

Campbell, J. E., and Carlson, M. 2002. "Panopticon. com: Online surveillance and the commodification of privacy," Journal of Broadcasting & Electronic Media (46:4), Taylor & Francis, pp. 586–606.

Cavoukian, A., and Emam, K. El. 2011.*Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy*, Information and Privacy Commissioner of Ontario, Canada.

Chaudhuri, K., and Monteleoni, C. 2009. "Privacy-preserving logistic regression," in *Advances in Neural Information Processing Systems*, pp. 289–296.

Chellappa, R. K., and Sin, R. G. 2005. "Personalization versus privacy: An empirical examination of the online consumer's dilemma," Information Technology and Management (6:2-3), Springer, pp. 181–202.

Chen, B.C., Kifer, D., LeFevre, K., and Machanavajjhala, A. 2009. "Privacy-Preserving Data Publishing," *Foundations and Trends in Databases* (2:1–2)Now Publishers Inc., pp. 1–167.

Chen, H., Chiang, R., and Storey, V. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact.," *MIS Quarterly* (36:4), pp. 1165–1188.

Czepiel, S. A. 2002. "Maximum likelihood estimation of logistic regression models: theory and implementation," *Available at czep.net/stat/mlelr.pdf* .

Davies, S. G. 1997. "Re-engineering the right to privacy: how privacy has been transformed from a right to a commodity," in Technology and privacy, pp. 143–165.

Dinev, T., and Hart, P. 2006. "An extended privacy calculus model for e-commerce transactions," Information Systems Research (17:1), INFORMS, pp. 61–80.

Du, W., Han, Y. S., and Chen, S. 2004. "Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification.," in *SDM* (Vol. 4), pp. 222–233.

Duncan, G., and Stokes, L. 2009. 'Data masking for disclosure limitation,' Wiley Interdisciplinary Reviews: Computational Statistics (1:1), pp. 83–92.," *Wiley Interdisciplinary Reviews: Computational Statistics* (1:1), pp. 83–92.

Duncan, G. T., Elliot, M., and Salazar-González, J.-J. 2011. *Statistical Confidentiality*, New York, NY: Springer New York.

Gentry, C. 2009. "Fully Homomorphic Encryption Using Ideal Lattices," in STOC, (Vol. 9) , pp. 169–178.

Goldfarb, A., and Tucker, C. 2012. "Shifts in privacy concerns," *The American Economic Review* (102:3)American Economic Association, pp. 349–353.

Goldfarb, A., and Tucker, C. E. 2011. "Privacy Regulation and Online Advertising," Management Science (57:1), pp. 57–71 (doi: 10.1287/mnsc.1100.1246).

Guo, C.H., and Higham, N. J. 2006. "A schur-newton method for the matrix pth root and its inverse," *SIAM journal on matrix analysis and applications* (28:3)Society for Industrial and Applied Mathematics, pp. 788–804.

Hall, R., Fienberg, S., and Nardi, Y. 2011. "Secure multiple linear regression based on homomorphic encryption," *Journal of Official Statistics* , pp. 1–23.

Hann, I.H., Hui, K.-L., Lee, S.-Y.T., and Png, I. P. L. 2007. "Overcoming online information privacy concerns: An information-processing theory approach," *Journal of Management Information Systems* (24:2)ME Sharpe, pp. 13–42.

Hevner, A., and Chatterjee, S. 2010. *Design Research in Information Systems: Theory and Practice (Google eBook)*, Springer, p. 348.

Hevner, A., March, S., Park, J., and Ram, S. 2004. *Design Science in Information Systems Research*, *Management Information Systems Quarterly*, (Vol. 28) .

Hui, K.-L., Tan, B. C. Y., and Goh, C.-Y. 2006. "Online information disclosure: Motivators and measurements," ACM Transactions on Internet Technology (TOIT) (6:4), ACM, pp. 415–441.

Information Shield. 2011. "International Privacy Laws," *International Data Privacy Laws*, .

Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. 2005. "Secure Regression on Distributed Databases," *Journal of Computational and Graphical Statistics* (14:2)Taylor& Francis, pp. 263–279.

Klopfer, P. H., and Rubenstein, D. I. 1977. "The concept privacy and its biological basis," Journal of social Issues (33:3), Wiley Online Library, pp. 52–65.

Krasnova, H., and Veltri, N. F. 2010. "Privacy Calculus on Social Networking Sites: Explorative Evidence from Germany and USA," in 2010 43rd Hawaii International Conference on System Sciences, IEEE, pp. 1–10 (doi: 10.1109/HICSS.2010.307).

Laudon, K. C. 1996. "Markets and privacy," Communications of the ACM (39:9), ACM, pp. 92–104.

Laufer, R. S., and Wolfe, M. 1977. "Privacy as a concept and a social issue: A multidimensional developmental theory," Journal of Social Issues (33:3), Wiley Online Library, pp. 22–42.

Li, T., and Li, N. 2009. "On the tradeoff between privacy and utility in data publishing," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, New York, New York, USA: ACM Press, June 28, p. 517 (doi: 10.1145/1557019.1557079).

Narayanan, A., and Shmatikov, V. 2008. "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, , May, pp. 111–125.

Ohm, P. 2010. "Broken Promises Of Privacy: Responding To The Surprising Failure Of Anonymization. ," *UCLA Law Review. Aug2010* (57:6), pp. 1701–1777.

Paillier, P. 1999. "Public-key cryptosystems based on composite degree residuosity classes," Springer-Verlag, pp. 223–238.

Posner, R. A. 1981. "The economics of privacy," The American economic review (71:2), JSTOR, pp. 405–409.

Purewal, S. 2014. "Loaded and Locked: 3 Seriously Secure Cloud Storage Services." PCWorld. http://www.pcworld.com/article/2105100/loaded-and-locked-3-seriously-secure-cloud-storage-services.html (accessed September 26, 2014).

Rastogi, V., Suciu, D., and Hong, S. 2007. "The boundary between privacy and utility in data publishing," VLDB Endowment, pp. 531–542 (available at http://dl.acm.org/citation.cfm?id=1325851.1325913).

Rivest, R. L., Adleman, L., and Dertouzos, M. L. 1978. "On data banks and privacy homomorphisms," *Foundations of secure computation* (4:11), pp. 169–180.

Rosen, J. 2012. "The Right to Be Forgotten," Stanford Law Review Online (64), p. 88 (available at http://www.stanfordlawreview.org/online/privacy-paradox/right-to-be-forgotten?em_x=22).

Sankar, L., Rajagopalan, S. R., and Poor, H. V. 2013. "Utility-privacy tradeoffs in databases: An information-theoretic approach," Information Forensics and Security, IEEE Transactions on (8:6), IEEE, pp. 838–852.

Sheth, J. N. 2011. "Impact of emerging markets on marketing: Rethinking existing perspectives and practices," *Journal of Marketing* (75:4)American Marketing Association, pp. 166–182.

Siponen, M. T. 2005. "Analysis of modern IS security development approaches: towards the next generation of social and adaptable ISS methods," *Information and Organization* (15:4), pp. 339–375 (doi: 10.1016/j.infoandorg.2004.11.001).

Smith, R. E. 2002. *Compilation of State and Federal Privacy Laws*, Privacy Journal, p. 103.

Smith, H. J., Dinev, T., and Xu, H. 2011. "Information privacy research: an interdisciplinary review," *MIS Quarterly* (35:4), Society for Information Management and The Management Information Systems Research Center, pp. 989–1016 (available at http://dl.acm.org/citation.cfm?id=2208940.2208950).

Solove, D. J. 2006. "A taxonomy of privacy," *University of Pennsylvania Law Review* (154), JSTOR, pp. 477–564.

Stone, E. F., and Stone, D. L. 1990. "Privacy in organizations: Theoretical issues, research findings, and protection mechanisms," Research in personnel and human resources management (8:3), Sage Publications, pp. 349–411.

Sweeney, L. 2000. Simple Demographics Often Identify People Uniquely, Institute for Software Research. *Working Paper*.

Tucker, C. 2014. "Social Networks, Personalized Advertising and Privacy Controls," Journal of Marketing Research (51:5), American Marketing Association, pp. 546–562 (doi: 10.1509/jmr.10.0355).

Warren, S. D., and Brandeis, L. D. 1890. "The right to privacy," Harvard law review (4:5), pp. 193–220.

Westin, A. F. 1968. "Privacy and freedom," *Washington and Lee Law Review* (25:1), p. 166.

Wilson, D., and Ateniese, G. 2014. "To Share or Not to Share in Client-Side Encrypted Clouds," *arXiv preprint arXiv:1404.2697*.

Xu, H., Teo, H.-H., Tan, B. C. Y., and Agarwal, R. 2012. "Research Note-Effects of Individual Self-Protection, Industry Self-Regulation, and Government Regulation on Privacy Concerns: A Study of Location-Based Services," *Information Systems Research* (23:4)INFORMS, pp. 1342–1363.

Xu, H., Teo, H.-H., Tan, B. C. Y., and Agarwal, R. 2009. "The role of push-pull technology in privacy calculus: the case of location-based services," Journal of Management Information Systems (26:3), ME Sharpe, pp. 135–174.

Yu, M., and Liu, P. 2007. "Database isolation and filtering against data corruption attacks," in *Twenty-Third Annual Computer Security Applications Conference, 2007. ACSAC 2007,* pp. 97–106.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., and Winslett, M. 2012. "Functional mechanism: regression analysis under differential privacy," *Proceedings of the VLDB Endowment* (5:11), VLDB Endowment, pp. 1364–1375.