ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

K-Nearest Neighbor Classification Mechanism of Secure Encrypted Relational Data

Popat Borse[1], Akshay Dabi[2], Vivek Thorat[3], Pranay Bamane[4], Arslan Shaikh[5]

Professor, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering Lohagaon, Pune, Savitribai Phule Pune University, Pune, India[1]

Student, Dept. of Computer Engineering, Dr. D. Y. Patil School of Engineering Lohagaon, Pune, Savitribai Phule Pune University, Pune, India[2,3,4,5]

ABSTRACT: Information Mining has wide applications in numerous ranges, for example, managing an account, pharmaceutical, investigative exploration and among government offices. Order is one of the ordinarily utilized undertakings as a part of information mining applications. For as long as decade, because of the ascent of different protection issues, numerous hypothetical and commonsense answers for the grouping issue have been proposed under various security models. Be that as it may, with the late notoriety of distributed computing, clients now have the chance to outsource their information, in encoded structure, and also the information mining errands to the cloud. Since the information on the cloud is in encoded structure, existing protection safeguarding grouping systems are not pertinent. In this paper, we concentrate on tackling the characterization issue over scrambled information. Specifically, we propose a protected k-NN classifier over encoded information in the cloud. The proposed k-NN convention secures the classification of the information, client's information question, and information access designs. To the best of our insight, our work is the first to build up a safe k-NN classifier over encoded information under the standard semi-legit model. Likewise, we experimentally break down the productivity of our answer through different analyses.

KEYWORDS: Security, k-NN Classifier, Outsourced Databases, Encryption, privacy preserving.

I. INTRODUCTION

As of late, the distributed computing worldview is upsetting the associations' method for working their information especially in the way they store, get to and handle information. As a developing figuring worldview, distributed computing pulls in numerous associations to consider genuinely with respect to cloud potential as far as its cost-productivity, adaptability, and offload of managerial overhead. Frequently, associations appoint their computational operations notwithstanding their information to the cloud. In spite of huge points of interest that the cloud offers, protection and security issues in the cloud are anticipating organizations to use those favorable circumstances. At the point when information are profoundly delicate, the information should be scrambled before outsourcing to the cloud. Be that as it may, when information are scrambled, regardless of the basic encryption plan, performing any information mining assignments turns out to be exceptionally testing while never decoding the information. The information proprietor outsources his/her database and DBMS functionalities (e.g., kNN inquiry) to an untrusted outer administration supplier which deals with the information in the interest of the information proprietor where just trusted clients are permitted to question the facilitated information at the administration supplier. By outsourcing information to an untrusted server, numerous security issues emerge, for example, information security (shielding the privacy of the information from the server and in addition from inquiry guarantor). To accomplish information security, information proprietor is required to utilize information Anonymization models (e.g., k-namelessness) or cryptographic (e.g., encryption and information annoyance) methods over his/her information before outsourcing them to the server. Encryption is a conventional procedure used to ensure the classification of delicate information, for example, therapeutic records. Because of information encryption, the procedure of inquiry assessment over scrambled information gets to be testing. Along this bearing, different procedures have been proposed for preparing extent and total inquiries over encoded information. Utilizing encryption as an approach to accomplish information privacy may

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

bring about another issue amid the inquiry handling venture in the cloud. When all is said in done, it is extremely hard to prepare scrambled information while never decrypting it. The inquiry here is the way the cloud can execute the questions over scrambled information while the information put away at the cloud are encoded at all times.

II. GOALS AND OBJECTIVE

1. Improving the efficiency of SMINn is an important first step for improving the performance of our PPkNN protocol.
2. Our protocol protects the confidentiality of the data, user's input query, and hides the data access patterns.
3. We also evaluated the performance of our protocol under different parameter settings.

III. MOTIVATION OF THE PROJECT

We propelled the PPKNN over scrambled information to accomplish economies of scale for Cloud Computing. At that point we presented new security primitives, specifically secure least (SMIN), secure recurrence (SF), and proposed new answers for them. Second, the work in did not give any formal security examination of the hidden sub-conventions. Then again, this paper gives formal security evidences of the basic sub-conventions and in addition the PPkNN convention under the semi-legitimate model. We demonstrate that our proposed arrangement is secure and protection saving, while accurately understanding the objective of PPKNN. In [1] it presented another down to earth system for remote information stockpiling with productive access design security and accuracy. A customer can send this procedure to issue scrambled perusing, composing, and embeds to a very malignant stockpiling administration suppliers, without giving data or access designs. The supplier can't build up any connection between's later getting to, or to recognize a read and a compose. Additionally, the customer is given with solid rightness certifications to its operations supplier conduct does not go undetected. The fabricated functional framework as requests of greatness quicker than existing usage that can perform over numerous inquiries every second on 1.5TB+ databases with full calculation security and accuracy. The proposed completely homomorphism encryption plans in [2] a plan that permits anybody to utilize circuits over scrambled information without having the capacity to decode. The arrangement comes in 3 stages; initial step was to give some broad result that, to make an encryption plan that licenses assessment of a circuit. Its own decoding circuit we call a plan that can take results from its unscrambling circuit's boots trappable. Cross section based cryptography normally have unscrambling calculations with low level circuits appearances, frequently higher by an inside item calculation that is in NC1. The perfect cross sections give both added substance and multiplicative homo morphisms, as expected to assess general circuits. Tragically, the underlying plan is not exactly boots trappable i.e., the profundity of the plan can accurately choose whether it will be logarithm in the cross section measurement, much the same as the profundity of the unscrambling circuit, however the past is more prominent than this one. In the last step, it demonstrate to change the plan to decrease the profundity of the unscrambling circuit, and in this way acquire a boots trappable encryption plan, without lessening the profundity that the plan can assess. To put it plainly, we finish this by enacting the encoded to begin the decoding Process, leaving less work for the decoded, as like the server leaves less work for the unscrambled in a server supported cryptography based framework. As proposed in [4] completely homomorphism framework can take care of the issue as outsider is utilized for the subjective capacities however such systems are unreasonable. By utilizing Shamir's plan [3] we can create PPkNN yet our work is diverse in other way. Existing work on PP Data Mining (either irritation or secure multi-party calculation based methodology) can't take care of the DMED issue. Uneasy information don't have semantic security, so information bother strategies can't be utilized to encode exceedingly fragile information. Likewise the uneasy information doesn't create exceptionally precise information mining results.

IV. EXISTING SYSTEM PROBLEM

Suppose Alice owns a database D of n records t1. . . tnand m + 1 attributes. Let ti,jdenote the jthattribute value of record ti. Initially, Alice encrypts her database attribute-wise, that is, she computes Epk(ti,j), for $1 \leq i \leq n$ and $1 \leq j \leq m+1$, where column (m+1) contains the class labels. We assume that the underlying encryption scheme is semantically secure . Let the encrypted database be denoted by D′. We assume that Alice outsources D′ as well as the future

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

classification process to the cloud.Let Bob be an authorized user who wants to classify his input record q = hq1, . . . ,qmiby applying the k-Classification method based on D′. We refer to such a process as privacy-preserving k-NN (PPkNN) classification over encrypted data in the cloud. Formally, we define the PPkNN protocol as:
PPkNN(D′, q) → cqwhere cqdenotes the class label for q after applying k-NN classification method on D′ and q.

V. PROPOSED SYSTEM APPROACH

In proposed system various module are required for successful getting expected outcome at the different module levels. The system is first developed in small codes called units, which are integrated in the next phase with other units. All of the below protocols are considered under two-client semi-honest setting. In particular, we consider the presence of two semi honest clients P1 and P2 such that the Pallier's secret key sk is known only to P2 whereas ik is publicSecure Minimum (SMIN):- In this protocol, P1holds private input (u′, v′) and P2 holds sk, where u′ = ([u],Epk(su)) and v′ = ([v],Eik(sv)). Here su (resp., sv) denotes the secret associated with u(resp., v). The goal of SMIN is for P1 and P2 to jointly Here we present a set of generic sub-protocols that will be used in constructing our proposed k-NN protocol .All of the below protocols are considered under two-clients semi-honest setting. In particular, we consider the presence of two semi honest clients P1 and P2 such that the Palliser's secret key sk is known only to P2 whereas ik is public.

Secure Minimum out of n Numbers (SMINn):- In this protocol, we consider P1 with n encrypted vector's ([d1], [dn]) along with their corresponding encrypted secrets and P2 with sk. Here [dp] = Eik(dp,1), . . . ,Eik(dp,l)i where dp,1 and di,1 are the most and least significant bits of integer irrespectively, for $1 \leq p \leq n$. The secret ordp is given by sdi . P1 and P2 jointly compute [min(d1, . . . , dn)]. In addition, they compute Epk(smin(d1,...,dn)). At the last of this protocol ,the output ([min(d1, . . . , dn)],Epk(smin(d1,...,dn)))is known only to P1. During SMINn, no information regarding any of dp's and their secrets is revealed to P1 and P2..Secure Frequency (SF):- Here P1 with private input (hEik(c1), . . .Eik(cw)p, hEik(c′1), . . . ,Eik(c′k)p)and P2 securely compute the encryption of the frequency of cq , denoted by f(cq), in the listhc′1, . . . , c′kp, for $1 \leq q \leq w$. Here we explicitly assume that cq's are unique and $c'p \in \{c1, \ldots, cw\}$, for $1 \leq p \leq k$. The output Eik(f(c1)), . . . ,Eik(f(cw))p will be known only to P1. During the SF protocol, no data regarding c′p, cq , and f(cq) is revealed to P1 andP2, for $1 \leq p \leq k$ and $1 \leq q \leq w$.

Secure Computation of Majority Class (SCMCk):- Without loss of generality, let us assume that Sam's dataset D consists of x unique class labels denoted by c = hc1, . . . ,cwi. We assume that Sam outsources his list of encrypted classes to C1. That is, Alice outsources hEpk(c1), . . . ,Epk(cw)i to C1 along with here encrypted database D during the data outsourcing step. Note that, for security reasons, Sam may add test categories into the list to protect the total number of class labels, i.e. x from C1 and C2. However, we assume that Sam does not add any categories to c.

VI. LITERATURE SURVEY

1. **Project Title: Survey on Privacy Preserving Data Mining**

From this paper We Referred

Data mining is the extraction of interesting patterns or knowledge from huge amount of data. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. A number of methods and techniques have been developed for privacy preserving data mining. This paper provides a wide survey of different privacy preserving data mining algorithms and analyses the representative techniques for privacy preserving data mining, and points out their merits and demerits. Finally the present problems and directions for future research are discussed.

2. **Project Title: Proving in Zero-Knowledge that a Number Is the Product of Two Safe Primes.**

From this paper We Referred

We present the reticent statistical zero-knowledge protocols to prove statements such as:

- A committed number is a prime.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

- A committed (or revealed) number is the product of two safe primes, i.e., primes p and q such that (p − 1) =2and (q − 1) =2areprime.
- A given integer has large multiplicative order modulo a composite number that consists of two safe prime factors.

3. Project Title: Secure k-Nearest Neighbor Query over Encrypted Data in Outsourced Environments.
From this paper We Referred

For the past decade, query processing on relational data has been studied extensively, and many theoretical and practical solutions to query processing have been proposed under various scenarios. With the recent popularity of cloud computing, users now have the opportunity to outsource their data as well as the data management tasks to the cloud. However, due to the rise of various privacy issues, sensitive data (e.g., medical records) need to be encrypted before outsourcing to the cloud. In addition, query processing tasks should be handled by the cloud; otherwise, there would be no point to outsource the data at the first place. To process queries over encrypted data without the cloud ever decrypting the data is a very challenging task. In this paper, we focus on solving the k-nearest neighbor (kNN) query problem over encrypted database outsourced to a cloud: a user issues an encrypted query record to the cloud, and the cloud returns the k closest records to the user. We first present a basic scheme and demonstrate that such a naive solution is not secure. To provide better security, we propose a secure kNN protocol that protects the confidentiality of the data, user's input query, and data access patterns. Also, we empirically analyze the efficiency of our protocols through various experiments. These results indicate that our secure protocol is very efficient on the user end, and this lightweight scheme allows a user to use any mobile device to perform the kNN query.

4. Project Title: Managing and Accessing Data in the Cloud Privacy Risks and Approaches
From this paper We Referred
Ensuring proper privacy and protection of the information stored, communicated, processed, and disseminated in the cloud as well as of the users accessing such information is one of the grand challenges of our modern society. As a matter of fact, the advancements in the Information Technology and the diffusion of novel paradigms such as data outsourcing and cloud computing, while allowing users and companies to easily access high quality applications and services, introduce novel privacy risks of improper information disclosure and dissemination. In this paper, we will characterize different aspects of the privacy problem in emerging scenarios. We will illustrate risks, solutions, and open problems related to ensuring privacy of users accessing services or resources in the cloud, sensitive information stored at external parties, and accesses to such information.

5. Project Title: Privacy-preserving data mining in the malicious model
From this paper We Referred

Most of the cryptographic work in privacy-preserving distributed data mining deals with semi-honest adversaries, which are assumed to follow the prescribed protocol but try to infer private information using the messages they receive during the protocol. Although the semi-honest model is reasonable in some cases, it is unrealistic to assume that adversaries will always follow the protocols exactly. In particular, malicious adversaries could deviate arbitrarily from their prescribed protocols. Secure protocols that are developed against malicious adversaries require utilization of complex techniques. Clearly, protocols that can withstand malicious adversaries provide more security. However, there is an obvious trade-off: protocols that are secure against malicious adversaries are generally more expensive than those secure against semi-honest adversaries only. In this paper, our goal is to make an analysis of trade-offs between performance and security in privacy-preserving distributed data mining algorithms in the two models. In order to make a realistic comparison, we enhance commonly used sub protocols that are secure in the semi-honest model with zero knowledge proofs to be secure in the malicious model. We compare the performance of these protocols in both models.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

VII. METHODOLOGY USED/ SYSTEM ARCHITECTURE

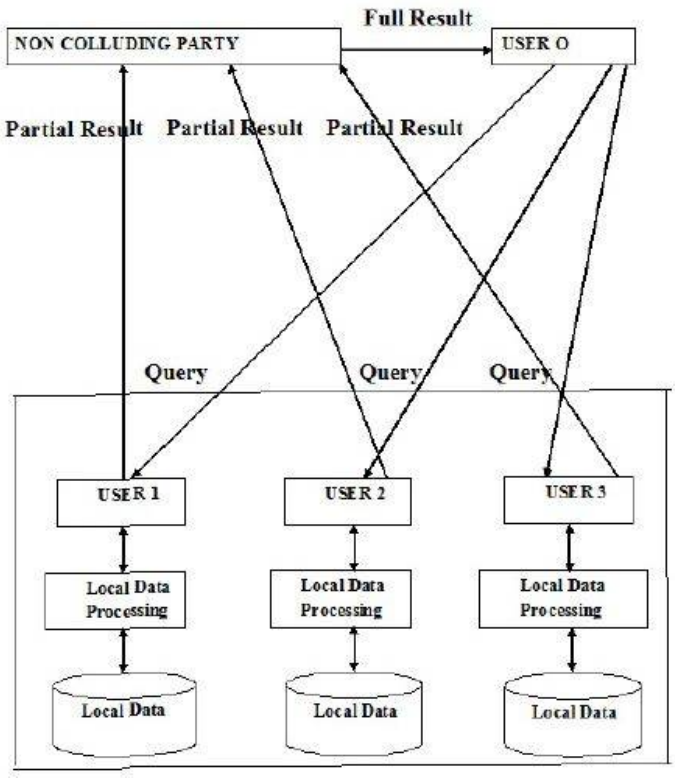

Fig No 01 System Architecture

1. Data confidentiality - Contents of T or any intermediate results should not be revealed to the cloud.
2. Query privacy - Bob's input query Q should not be revealed to the cloud.
3. Correctness - The output (t'1……t'k) should be revealed only to Bob. In addition, no information other thant'1…..t'kshould be revealed to Bob.
4. Low computation overhead on Bob - After sending his encrypted query record to the cloud, Bob involves only in a little computation compared with the existing works. More details are given in Section.
5. Hidden data access patterns - Access patterns to the data, such as the records corresponding to the k-nearest neighbors of Q, should not be revealed to Alice and the cloud (to prevent any inference attacks).

VIII. ALGORITHM AND TECHNIQUE USED

1. Search algorithm.

The search process of our DMRS scheme starts from the root node with a recursive procedure upon the tree in a special depth-first manner, which is called as "Greedy Depth first Traverse Strategy". Specifically, if the node's similarity score is less than or equal to the minimum similarity score of the currently selected top-k documents, search process returns to the parent node, otherwise, it goes down to examine the child node. The similarity score of each node u is calculated as Formula (1), i.e., the inner product of query vector Q and data vectorDu . This procedure is executed recursively until the objects with top- k scores are selected. The search can be done very efficiently, since only part of the index tree is visited due to the relatively accurate maximum score prediction. Algorithm 1 shows the process of our proposed search scheme.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

2. **Secure algorithm.**

As soon as the plaintext index tree is built, secure encryption scheme needs to be executedto prevent information leakage. We adopt the encryption scheme in [4] to secure our indextree, the whole process is described as follows:

1. Setup
2. GenIndex
3. GenTrapdoor
4. SimEvaluation

VIII. APPLICATIONS

It is used in cloud when we store the data on cloud in encrypted format. And access the data using secret key in decrypted form. We know the k-NN classifier and developed a privacy-preserving protocol for it over encrypted data.

1. Increased website ROI

When users can find what they are looking for on a website easily they are more likely to take the desired action, whether that be a product purchase, information request, or simply learn what they wanted to know.

2. Reduced customer service costs

Providing a self-service means to access common information on a website can reduce the number of calls or emails to customer service. In addition customer service can use the same search when answering questions.

3. Increased productivity

If you are like many companies you have file shares full of documents, but aren't sure exactly what is there or where it is. By providing a quality search you can quickly locate both the documents you are looking for, as well as related documents that may already exist, preventing duplicate effort.

IX. EXPERIMENTAL SETUP AND RESULT

Here discussed some experiments demonstrating the performance of Privacy Preserving k-Nearest Neighbor (PPKNN) classification method with some parameter settings. The Partial Homomorphic encryption scheme is used as the underlying additive homomorphic encryption scheme and implemented the proposed PPkNN protocol in JAVA.

1 Dataset Details and Experimental Setup

This protocol used the Car Evaluation dataset from the UCI KDD archive[12]. This dataset consists of 1728 records (n = 1728) and 6 attributes (m = 6). There is a separate class attribute and the dataset is categorized into four different classes (w = 4). Encrypted this dataset attribute wise, using the Homomorphic encryption the key size is varied in experiments, and the encrypted data were stored on server machine. Based on PPkNN protocol, executed a randomquery over this encrypted data.

2 Performance of Privacy Preserving k-Nearest Neighbor Classification-

Partial Homomorphic Encryption The encryption key size K is either 512 or 1024 bits if is K=512bits, the computation cost varies from 9.98 to 46.16 minutes when k is changed from 5 to 25, respectively. When K=1024 bits, the computation cost varies from 66.97 to 309.98 minutes when k is changed from 5 to 25, respectively. For K=512 bits, the computation time for Stage 2 to generate the final class label corresponding to the input query varies from 0.118 to 0.285 seconds when k is changed from 5 to 25. For K=1024 bits, Stage 2 took 0.789 and 1.89 seconds when k = 5 and 25, respectively. Here observed that the computation time of Stage 1 accounts for at least 99 percentage of the total time in PPkNN. For example, when k = 10 and K=512bits, the processing costs of Stage 1 and 2 are 19.06 minutes and

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

0.175 seconds, respectively. Under this scenario, cost of Stage 1 is 99.98 percentage of the total cost ofPPkNN. The total computation time of PPkNN grows almost linearly with n and k.

X. RESULTS

As mentioned in Section 2.3, to formally prove that SMIN is secure under the semi-honest model, we need to show that the simulated image of SMIN is computationally indistinguishable from the actual execution image of SMIN. An execution image generally includes the messages exchanged and the information computed from these messages.

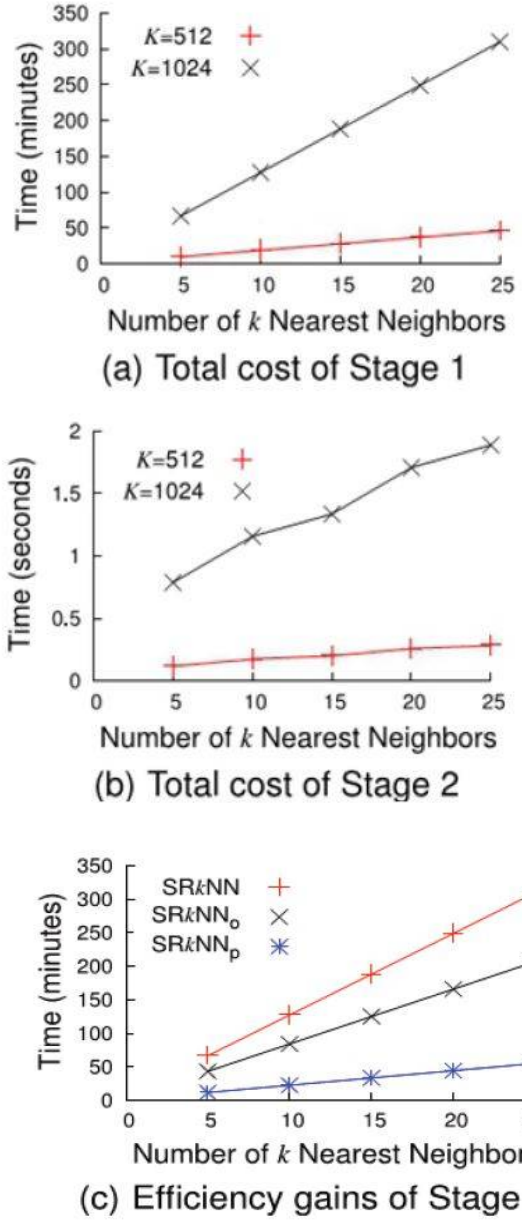

Fig No 02 Computation costs of PPkNN for varying number of k nearest neighbors and encryption key size K

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

X. CONCLUSION

To ensure client security, different protection saving classification strategies has been proposed over the previous decade. The current procedures are not appropriate to out sourced database environment s where the information lives in scrambled structure on an outsider server. This paper proposed a novel security safeguarding k-NN classification convention over encoded information in the cloud. Our convention secures the confidentiality of the information, client's info inquiry, and shrouds the information get to pat terns. We additionally assessed the execution of our convention under various parameter set tings. Since enhancing the efficiency of SMIN n is an import subterranean insect first step f or enhancing the - execution of our PPkNN convention, we plan to explore option and more efficient solute-particles t o t he SMIN n issue in our future work. Additionally, we will examine and extend our exploration to other classification calculations.

REFERENCES

[1] Mell and T. Grance, "The nist definition of cloud computing(draft)," NIST special publication , vol. 800, p. 145, 2011.
[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "Managing and accessing data in the cloud: Privacy risks and approaches," in CRiSIS, pp. 1 –9, 2012.
[3] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud: practical access pattern privacy and correctness on untrusted storage," in ACM CCS , pp. 139–148, 2008.
[4] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in Eurocrypt , pp. 223–238, 1999.
[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted re-lational data." eprint arXiv:1403.5001, 2014.
[6] C. Gentry, "Fully homomorphic encryption using ideal lattices," in ACM STOC , pp. 169–178, 2009.
[7] C. Gentry and S. Halevi, "Implementing gentry's fully- homomorphic encryption scheme," in EUROCRYPT , pp. 129– 148, Springer, 2011.
[8] A. Shamir, "How to share a secret,"Commun.ACM, vol. 22, pp. 612–613, 1979.
[9] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations,"in Proc. 13th Eur.Symp. Res. Comput. Security: Comput.Security, 2008, pp. 192–206.
[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining,"ACMSigmod Rec., vol. 29, pp. 439–450, 2000.
[11]P. Williams, R. Sion, and B. Carbunar, Building castles out of mud: practical access pattern privacy and correctness onuntrusted storage, in ACMCCS, pp. 139148, 2008.
[12]C. Gentry, Fully homomorphic encryption using ideal lattices,in ACM STOC, pp. 169178, 2009.
[13]Shamir, How to share a secret, Commun. ACM, vol. 22,pp. 612613,B. K. Samanthula, Y. Elmehdwi, and W. Jiang, k-nearest neighbor classification over semantically secure encrypted relationaldata. eprint arXiv:1403.5001,2014
[14]P. Paillier, "Public key cryptosystems based on compositedegreeresiduosity classes," in Eurocrypt, pp. 223–238, 1999.