Secure Efficient Encrypted Data Search Using Skyline Queries

B.SHAUMIYA [1] , DR.G.ARAVIND SWAMINATHAN [2]

[1]PG Student, [2]Professor

[1, 2,] Department of CSE, Francis Xavier Engineering College

ABSTRACT

The cloud server for outsourcing data and computation that provides a cost-effective way which supports large scale data storage and query processing. However, the medical records are particularly sensitive data that need to be protected from the cloud server and other unauthorized users due to security and privacy concerns. It supports various queries over encrypted data that remains a challenging task such that the cloud server does not gain any knowledge about the data, query, and query result in a secure and efficient way.Here,study about the problem of secure skyline queries over encrypted data. The skyline query is most importantly used for multi-criteria decision making by using kNN queries but also presents significant challenges due to its complex computations. Proposing a fully secure skyline query protocol on data encrypted using semantically-secure encryption. The new secure dominance protocol which act as a new key subroutine which can be also used as a building block for other queries.

Keywords: skyline query, encryption, security.

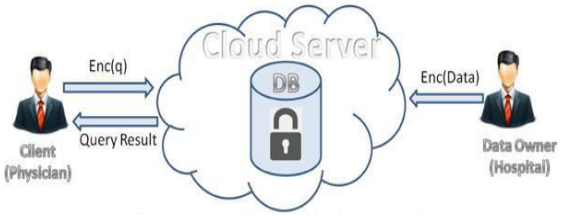
SKYLINE QUERIES

A skyline is defined as those points which are not dominated by any other point. A point dominates another point if it is as good or better in all dimensions and better in at least one dimension

1. INTRODUCTION

In a database, a skyline is a set of tuples of information (points) which stand out among the others because are of special interest to us. A common approach to protect the confidentiality of outsourced data is to encrypt the data. To protect the confidentiality of the query from cloud server, authorized clients also send encrypted queries to the cloud server. Below figure illustrates our problem

scenario of secure query processing over encrypted data in the cloud. The data owner outsources their encrypted data to the cloud server. The cloud server processes encrypted queries from the client on the encrypted data and returns the query result to the client During the query processing, the cloud server should not gain any knowledge about the data, data patterns, query, and query result.



CONTRIBUTION OF THIS WORK

- We study the secure skyline problem on encrypted data with semantic security for the first time. We assume the data is encrypted using the Paillier cryptosystem which provides semantic security and is partially homomorphic.
- We propose a fully secure dominance protocol, which can be used as a building block for skyline queries as well as other queries, e.g., reverse skyline queries and k-sky band queries.
- We present two secure skyline query protocols. The first one, served as a basic and efficient solution, leaks some indirect data patterns to the cloud server. The second one is fully secure and ensures that the cloud gains no knowledge about the data including indirect patterns. The proposed protocols exploit the partial (additive) homomorphism as well as novel permutation and perturbation techniques to ensure the correct result is computed while guaranteeing privacy.
- We provide security and complexity analysis of the proposed protocols. We also provide a complete implementation including both serial and parallelized versions which can be deployed in practical settings. We empirically study the efficiency and scalability of the implementations under different parameter settings, verifying the feasibility of our proposed solutions.

Consider a hospital who wishes to outsource its electronic health records to the cloud and the data is encrypted to ensure data confidentiality. Let P denote a sample heart disease dataset with attributes ID, age, trestbps (resting blood pressure). We sampled four patient records p1, …, p4 from the heart disease dataset of UCI machine learning repository as shown in Table I(a) and Figure 2. Consider a physician who is treating a heart disease patient q = (41, 125) and wishes to retrieve similar patients in order to enhance and personalize the treatment for patient q. While it is unclear how to define the

attribute weights for kNN queries (p1 is the nearest if only age is considered while p2, p3 are the nearest if only trestbps is considered), skyline provides all pareto-similar records that are not dominated by any other records. Given the query q, we can map the data points to a new space with q as the origin and the distance to q as the mapping function. The mapped records ti[j] = |pi[j] − q[j]| + q[j] on each dimension j are shown in Table I(b) and also in Figure 2. It is easy to see that t1 and t2 are skyline in the mapped space, which means p1 and p2 are skyline with respect to query q.

Our goal is for the cloud server to compute the skyline query given q on the encrypted data without revealing the data, the query q, the final result set {p1, p2}, as well as any intermediate result (e.g., t2 dominates t4) to the cloud. We note that skyline computation (with query point at the origin) is a special case of skyline queries. Our protocol can be also used

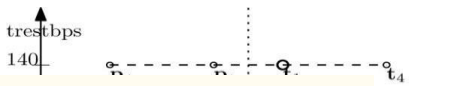


TABLE I

Sample of heart disease dataset.

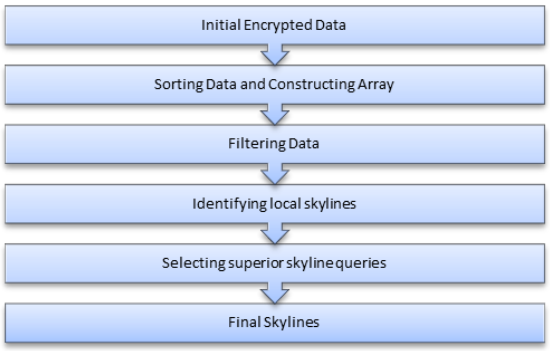
(a) Original data.		
ID	age	trestbps
p_1	40	140
p_2	39	120
p_3	45	130
p_4	37	140

(b) Mapped Data.		
ID	age	trestbps
t_1	42	140
t_2	43	130
t_3	45	130
t_4	45	140

for skyline computation.

II. SYSTEM FLOW DIAGRAM

This is our system flow diagram



III.METHODOLOGY

The following is our methodology of this work.

- Initial Encrypted Data
- Sorting data and constructing array
- Filtering data
- Identifying local skyline

– Retrieving local skyline

– Joining Skylines of all Relations

- Identifying Global Skylines

descriptions of this works are below

Initial Encrypted Data

- The first phase, identifying the skylines of each relation in all datacentres, attempts to identify the skylines of each relation separately, which are located at different datacenters, aiming at discarding all dominated data items from the join operation.
- Thereby it results in propagating only the most candidate data items into the next phases. This process assists by

avoiding joining of dominated data items via performing filtration.

- That leads in eliminating the unnecessary pairwise comparisons between data items and reduce the amount of data transfer significantly. The detail processes of this phase are elaborated in the following subsections.

Sorting data and constructing array

- This step is responsible for analysing the initial incomplete database relation and attempts to sort the data items based on non-missing dimensions in non-ascending order.
- Then a set of arrays is constructed and the id's of sorted data items are stored in connected arrays. The number of arrays constructed mainly depends on the number of dimensions with no-missing value.
- This step helps in reducing the searching space, which further leads to decrease the number of pairwise comparisons between data items in the subsequent phases.

Filtering data

- This step is one of the most significant phases in introducing the local skylines of each involved table. This phase is responsible for eliminating the dominated data items before applying skyline technique.
- This is achieved by scanning the whole data items in each array in sequential order using round robin fashion. The scanning process ends when all data items have been read at least once.
- It might happen that some data items are read more than once. Therefore, a counter is needed to count the number of reading of each data item.
- The idea behind using the counter is to sort the data items according to their count values in decreasing order. Hence, the data items with the highest count score have a higher potential to be in the skylines set.
- Besides, it also helps in eliminating a large number of dominated data items. The outputs of this process are a list of data items with their corresponding count values.

Identifying local skyline

- In this step, the data points that have no potential to be part of the skyline are eliminated before applying skyline technique.
- That also helps in reducing unnecessary pairwise comparisons to make the proposed approach more efficient.
- This eliminating process will be executed by removing all the data points from the list with count score less than

two. The rest data points will be stored in candidate set for the further process.

- This step is responsible for the implementation of skyline technique over the data items presented in the candidate set. The aim is to find the local skylines separately from all relations stored in different datacenters at distant locations. This process is conducted in parallel on all datacenters.
- That helps to reduce the maximum amount of data to be transferred from one data center to another for evaluation of final skylines. The process starts by reading the first data item in the candidate set and then compared with the remaining data items.
- The read data item named as processing data item p, while the data item to be compared with p is called candidate data item q. During the comparison process if p dominates q then q will be immediately eliminated from the candidate set.
- Else if neither p dominates q and nor q dominates p, then q will remain in the candidate set for further processing. However, if q dominates p then p will not be removed immediately; rather it will remain until the end of the iteration process.
- This is because p may have good potential to eliminate other data items and helps to sustain transitivity property and solves the issue of cyclic dominance. This process continues until all remaining data items are processed.
- It should be noted that no two data items are compared more than once. We argue that this process is effective in avoiding many unnecessary pairwise comparisons between data items. The output of this step is the set of the local skylines of each relation to be joined to form the final skylines.

Identifying Global Skylines

- This is the last phase of our proposed approach for processing skyline queries in a database with incomplete data over the cloud environment.
- It tries to determine the final skyline set which contains those data items that are not dominated by other data items in all involved relations.
- The sub-phases of the first phase (identifying the skylines of each relation) of our proposed approach will be performed on joined local skylines.
- If the joined data item is not dominated by the other data items in the candidate skyline set, then it is retrieved as part of the final skyline. Otherwise, it is removed from the candidate skyline set.

- In this process, we guarantee that the final skylines are the skylines of the relations in all cloud datacentres and no other data items might dominate the identified final skylines.

V.EXISTING SYSTEM

- The efficacy and efficiency of our schemes are thoroughly analysed and evaluated.
- Computational overhead and memory constraints become less prohibitive.
- The novel and efficient approach for computing the skyline in a secure multi-party computing environment is proposed without disclosing the individual attributes' value of the objects.
- The secure multi-party sorting protocol that uses the homomorphic encryption in the semi-honest adversary model is used for transforming each attribute value of the objects without changing their order on each attribute.
- Difficult in managing the password.
- For the first time, the concept of Reverse Skyline Queries is introduced. The multidimensional data set P is considered for the problem of dynamic skyline queries according to a query point q. This kind of dynamic skyline corresponds to the skyline of a transformed data space where point q becomes the origin and all points of P are represented by their distance vector to q.
- The reverse skyline query returns the objects whose dynamic skyline contains the query object q. In order to compute the reverse skyline of an arbitrary query point, the Branch and Bound algorithm (called BBRS)is proposed, which is an improved customization of the original BBS algorithm.

Disadvantages of existing system:

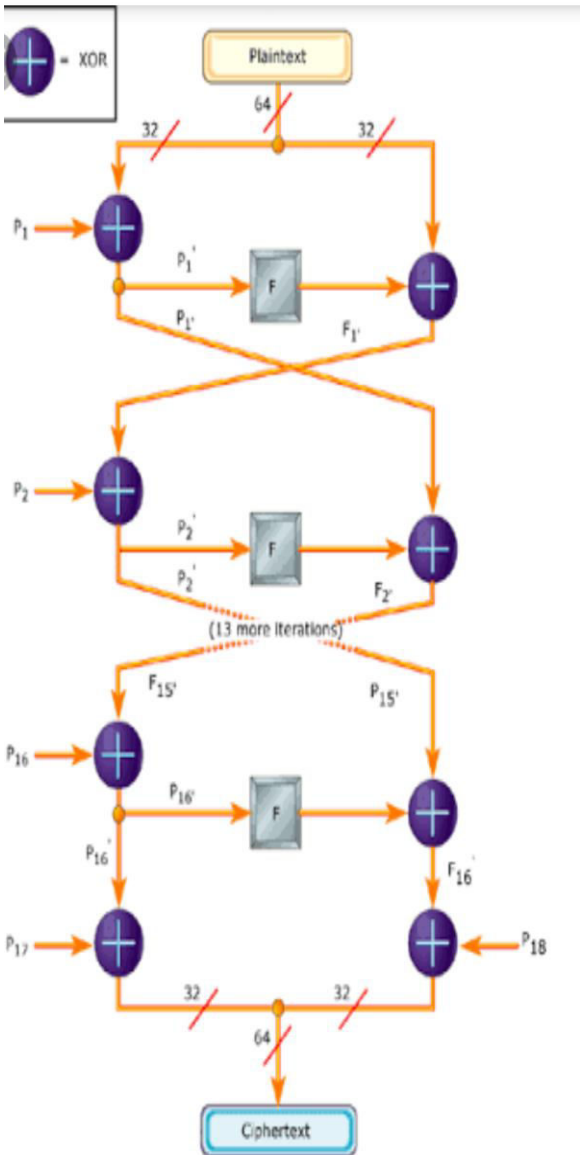
- Time consuming process.
- Not working for Cloud data.
- These results indicate that our secure protocol is very efficient on the user end, and this lightweight scheme allows a user to use any mobile device to perform the kNN query.
- Only semi supervised algorithm is used. Its accuracy is very low.

VI.PROPOSED SYSTEM

The secure skyline problem on encrypted data with semantic security for the first time is studied. The data is assumed to be encrypted using the Paillier cryptosystem which provides semantic security and is partially homomorphic. The fully secure dominance protocol is proposed, which can be used as a building block for skyline queries as well as other queries, e.g., reverse skyline queries and k-skyband queries. The two secure skyline query protocols is presented. The first one, served as a basic and efficient solution, leaks some indirect data patterns to the cloud server. The second one is fully secure and ensures that the cloud gains no knowledge about the data including indirect patterns. The proposed protocols exploit the partial (additive) homomorphism as well as novel permutation and perturbation techniques to ensure the correct result is computed while guaranteeing privacy. The security and complexity analysis of the proposed protocols is presented. The complete implementation including both serial and parallelized versions which can also be deployed in practical settings. Empirically, the study of the efficiency and scalability of the implementations under different parameter settings, verifying the feasibility of our proposed solutions

BLOWFISH ALGORITHM

Blowfish is a symmetric block cipher that can be used as a drop-in replacement for DES or IDEA.It takes a variable length key,from 32 bits to 448 bits,making it ideal for both domestic and exportable use.since it has been analyzed considerably,and it is slowly gaining acceptance as a strong encryption algorithm.Blowfish is also a cipher block meaning that it divides a message up into fixed length blocks during encryption and decryption.



VII.DATASET COLLECTIONS

This dataset is collected from following repository and web link

- Uci Repositary Heart Disease Dataset
- Records : 303 Nos
- Number Of Attributes : 14
- https://www.kaggle.com/ronitf/heart-disease-uci

VIII. CONCLUSION AND THE FURTHER WORK

In this work, we proposed a fully secure skyline protocol on encrypted data using two non-colluding cloud servers under the semi-honest model. It ensures semantic security in that the cloud servers knows nothing about the data including indirect data patterns, query, as well as the query result. This work has proposed a secure dominance sub-protocol. This work has proposed a fully secure skyline protocol. This work is demonstrated practical using simulation .Further optimization of algorithm complexity and running time.

IX.RESULTS

Blowfish AES Encryption Algorithm comparisons

Blowfish is a 16-round Feistel cipher and uses large key-dependent S-boxes. Blowfish is unpatented, license-free, and available free for all uses. Table 1 shows their characteristics. This Table Algorithms Characteristics

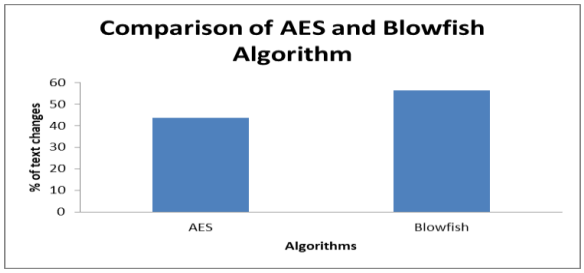
Factors	AES	Blowfish
Key Length	128 bits	448 bits
Block size	128 bits	64 bits
Speed	Slow	Fast
Security	Excellent Security	Secure Enough
Number of rounds	16	10

No of S-Boxes	4	1

In the Table above, a speed level comparison of the two symmetric algorithms has been done and based on the analysis we found that blowfish has the fastest speed compared with AES. Because of the largest key size the blowfish algorithm mentioned above has been given a fastest and strongest speed level compared with AES.

Integrity Checking in CBC mode

Algorithm	**Percentage changes in plain text**
AES	43.75 %
Blowfish	56.25 %



We have demonstrated that change in one bit in the key produces strong avalanche effect. The above graph means that Blowfish has a strong avalanche effect compared with others algorithms mentioned above.

X.REFERENCES

[1] W. Chen, M. Liu, R. Zhang, Y. Zhang, and S. Liu. Secure outsourced skyline query processing via untrusted cloud service providers. In INFOCOM 2016.

[2] Mahaboob,Qoasar,Asif Zamin,Annisa,. Privacy-Preserving Secure Computation of Skyline Query in Distributed Multi-Party Databases,2019.

[3] E. Dellis and B. Seeger. Efficient computation of reverse skyline queries. In VLDB, pages 291–302, 2007.

[4] S.Bothe,P.Karras,andA.Vlachou. eskyline:Processing skyline queries over encrypted data. PVLDB, 6(12):1338–1341, 2013.

[5] Y.Elmehdwi ,B.K.Samanthula,and W.Jiang. Securek-nearestneighbor query over encrypted data in outsourced environments. In ICDE 2014.

[6] W. K. Wong, D. W. Cheung, B. Kao, and N. Mamoulis. Secure knn computation on encrypted databases. In SIGMOD 2009.

[7] T. Veugen, F. Blom, S. J. A. de Hoogh, and Z. Erkin. Secure comparison protocols in the semi-honest model. J. Sel. Topics Signal Processing, 9(7):1217–1228, 2015.

[8] C. Y. Chan, H. V. Jagadish, K.-L. Tan, A. K. H. Tung, and Z. Zhang. Finding k-dominant skylines in high dimensional space. In SIGMOD Conference, pages 503–514, 2006.

[9] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game or A completeness theorem for protocols with honest majority. In ACM Symposium

on Theory of Computing, pages 218–229, 1987.

[10] V. Costan and S. Devadas. Intel sgx explained. Technical report, Cryptology ePrint Archive, Report 2016/086, 20 16. http://eprint. iacr.org.