IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 11, November 2016

Secure and Fast Sub-Graph Similarity Search in Outsourced Cloud Database with Data Deduplication

Nitin D.Dhamale[1], **Prof. Prakash P. Rokade** [2]

P.G. Student, Department of Computer Engineering, SND COE and RC, Yeola, India [1]

Professor & Head, Department of Information Technology, SND COE and RC, Yeola, India [2]

Abstract: Secure and Fast Sub-Graph search is system which is used over outsourced cloud database for search graph which are same to a queried by the client or user. Graph Data is increasing day by day, so database outsourcing is an solution to increasing graph database to the database owner. But, Cloud Database and query service Authority providers not trust or may be involved in attacks. In this paper, we propose authentication process techniques using Attribute base encryption for checking user credential either trusted or untrusted user to prevent tempering with cloud database. We propose a Fast sub graph retrieval technique using apriori algorithm .we also implement the Data deduplication to save the space on cloud using hashing generating function to avoid same data copy over cloud database. We also propose file ownership generation technique for owner of data. Our compressive results verify the effectiveness and efficiency of our proposed techniques.

Keywords: Sub-Graph Similarity Search, outsourced database, Data Deduplication.

I. INTRODUCTION

Graph is used to represent various complexes structure data in various filed like chemical industry a compound represented as graph ,biology a protein represented as graph, Web topology to design networks, social network of site ,attribute graph in computer vision. In such application sub graph searching is frequently used to search the graph. Example user gives query to graph database to find the related sub graph and often return result that match with to the user query (e.g.PubChem).

Similarity search is comes under the NP-hard problem. he most of owners of graph database not fully aware of the information technology resources and different techniques that is used for efficient search of their database. For example is user gives the query for alcohol structure to graph database which took 7 minutes to process query and desired result. This type of concert may not best for applications. Graph data base grow rapidly in volume recent study said that from 2006 to 2013 Pub hems database increase day by day from 58 GB to 142 GB. So it difficult to process huge amount of data with graph with an general service computers. The above mention reasons, outsourcing graph database are way to database owner .Dedicatedly to big data is delegate to a handling and controlling service provider (SP) which is third party. A client give query to service provider, as if he/she is access a utility and the Service provider provides query handling and process on behalf of data owner. Graph data storage outsourcing has been use by in lots of business. In lots of business. For example, in drug process engineering, lots business service providers sustain outsourcing of pharmacy databases. The service provided by the service provider may be not trusted. Service provider might possible involved in the attack there is possibility user may get receive tampered results. For example, Fig.1 shows an outsourced. Database with graph an molecular database D, a molecule query q and a threshold distance t 0:25.suppose that G4, G6 and G7 are graph result. This must be return as the queries result.

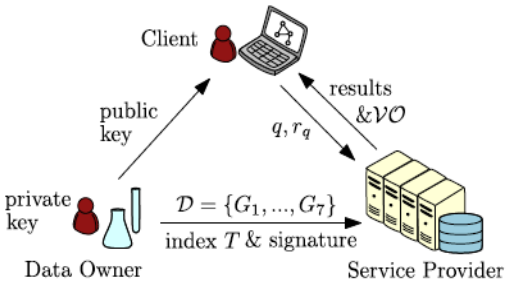

Fig. 1: Example of Outsourced graph database

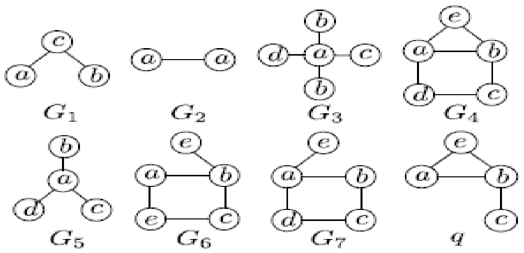

Figure 2: Query graph and Expected Result

DOI 10.17148/IJIREEICE.2016.41114

The service provider (SP) continuously consciously return incorrect result (e.g., G3) Fig.2, distort t to 0:1 or returns unfair result (e.g., only G4). This is limit the practicality of graph database outsourcing. so there is need of security and authentication mechanism is necessary over the database outsourcing . In most of place for graph searching filtering and verification framework is used, The filtering and verification used for indexing and filter the data, also used for the validate the queries result, But no study or any work done on authentication and encrypting, so in this paper more focus on authentication and Encoding of graph to reduce the time of uploading the graph over cloud compared to the existing system.

II. RELATED WORK

Horst Bunke [15] propose a new graph distance measure that is based on the maximal regular sub graph of two graphs. The main part of the paper is the formal verification that the new distance measure is a metric. An benefit of the new distance compute over graph edit distance is the fact that it does not depend on edit costs. It is well recognized that any edit distance measure significantly depends on the costs of the fundamental edit operations. But the difficulty how these edit costs are obtain is still unsolved. Using the new distance measure, this difficulty can be avoided.

Yuanyuan Zhu[16] discover the problem of finding top-k graphs in a graph database that are mainly similar to a query graph. This problem has been in many applications, like as image retrieving and chemical compound structure search. About the similarity measure in graph database, feature based and kernel based similarity measures have been used in the literature. But such measures are coarse and may lose the connectivity information among substructures

Dennis Shasha[17] there five distance-mapping algorithms and conduct experiment to compare their performance in data clustering applications. These comprise two algorithms called FastMap and MetricMap, and three hybrid heuristics that combine the two algorithms in dissimilar ways. Tentative results on both synthetic and RNA data show the superiority of the hybrid algorithms. The results involve that FastMap and MetricMap capture matching information about distance metrics and therefore can be used together to great advantage. The net outcome is that multi-day computations may be complete in minutes.

Xifeng Yan[18] investigate the issue of substructure similarity search using indexed features in outsource graph databases. By transform the edge recreation ratio of a query graph into the maximum allowed missing features, our structural filtering algorithms, called Grail, can filter many graphs without performing two of a kind wise similarity computations. It is further shown that using either too few or too a lot of features can result in poor filtering presentation. Thus the confront is to design an effective feature set selection strategy for filtering. By tentative the effect of different feature selection mechanism, we develop a multi-filter composition strategy, where each filter uses a separate and corresponding subset of the features.

III. PROBLEM STATEMENT

In this paper we proposed an Secure and Fast Sub-Graph Similarity search is system which is used over outsourced cloud database for search graph which are same to a queried by the client or user. Graph Data is increasing day by day, so database outsourcing is an solution to increasing graph database to the database owner. But, Cloud Database and query service Authority providers might not trust or may be involved in attacks. In this paper, we propose authentication process techniques using Attribute base encryption for checking user credential either trusted or untrusted user to prevent tempering with cloud database. we propose a Fast sub graph retrieval technique using apriori algorithm. In This paper, we propose authentication techniques using Attribute base encryption for checking user credential either trusted or untrusted user to prevent tempering with cloud database. we propose a Fast sub graph retrieval technique using apriori algorithm. So user get correct and fast result.

IV. PROPOSED WORK

We propose the system model, In which having four modules are authentication and credential checking ,cloud storage ,data owner ,query processing module. Data owner of graph upload the graph data over cloud storage in an encrypted form. The cloud storage performs mining of sub graph and also provides storage. The data owner or client give queries the cloud database for retrieval of similar sub graph. The authentication and credential checking of client and data owner is done by authentication and credential module .fig.3 show Architecture of proposed system

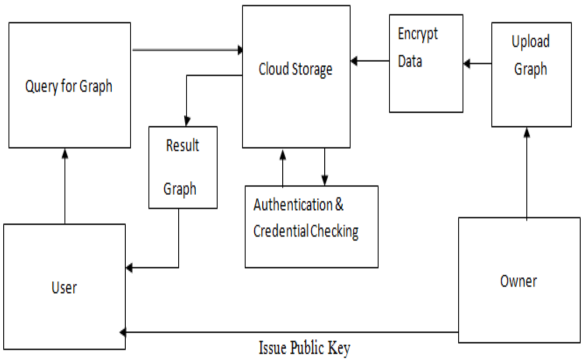

Fig. 3 Architecture of Proposed System

The above figure shows the Architecture of Proposed System.

DOI 10.17148/IJIREEICE.2016.41114

The system consists of following basic modules which are listed and explain below in detail.

A. Authentication and credential
This module is authenticating user either trusted or untrusted and check the user credential using attribute based encryption.

B. Cloud Storage
Outsourced Database which is used to store the user graph data and process the query.

C. Query Processing Module
This module is optimizing the user query to speed up the processing of query over cloud database also encrypt the query , mine the result from the cloud database using proposed apriori algorithm.

In prosed system we also implemented the data deduplication to avoid the same copy of data over cloud storage. In data deduplication the we calculate the hash value of data and store over the cloud if user try to upload same data file then the hash value is compared with hash table. If hash value is not matched then file is uploaded else messenger is shown to user file duplication has been found .we also embedded the secrete key with file for data owner ship. The user can use secrete key to challenge the owner ship of file.

D. Mathematical model:
Set Theory Let I be a set of Input to system and E is intermediate operation and D is set of output.
Input Set
I= {I1, I2, I3, I4.}
Where, I1=Graph Data. I2=User Name.
I3= Credential. I4= Query.
Intermediate Output Set.
E= {E1, E2, E3, E4}.
Where,
E1=Store Graph Data. E2= Authentication.
E3= Credential Checking. E4= Process Query.
Final Output Set.
D= {D1}.
Where, D1= Result Graph.
Following figure shows functional dependency of system:

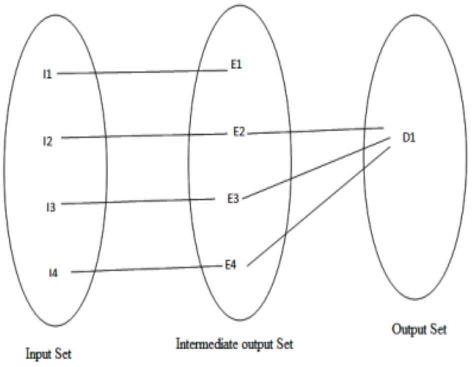

Fig 4: Functional Dependency of system

E. Algorithm
Apriori algorithm is proposed approach for the sub graph mining in outsource graph database.

Algorithm: AprioriGraph
Input
- D , A Graph Dataset;
- Min_sup, the minimum support threshold.

Output:
- S_k, the frequent substructure set

Method:

```
S_1– frequent single elements in the data set;
Call AprioriGraph (D, min_sup,S_1);
Procedure AprioriGraph (D, min_sup,S_1);
1: $S_{K+1} \longleftarrow Ø$;
2: For each frequent $g_1 \in S_K$ do
3: For each frequent $g_1 \in S_K$ do ∈
4: For each size (K+1) graph g formed by the
      merge of $g_i$ and   $g_j$ do
5: If g is frequent in D and $g \notin S_{K+1}$then
6: Insert g into $S_{K+1;}$
7: If $S_{K+1} \neq Ø$ then
8: AprioriGraph (D, min_sup, $S_{K+1}$);
9: return
```

V. RESULT AND DISCUSSION

The experiment with various dataset input set to the proposed system with respect to existing system proves that the proposed system having efficient encoding time. The time is in millisecond for encoding and Decoding.

TABLE I

Sr. No	Decoding (MS)	Encoding (MS)	Number of Scp
1	1	102	3
2	2	114	4
3	2	116	5
4	2	118	6
5	2	121	7
6	3	125	8

Form above Table I shows encoding & decoding time analysis when share size is K=2.

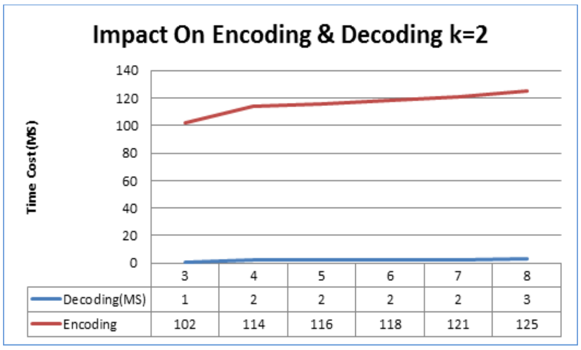

Fig. 3: Impact on Encoding/Decoding time: case 1 (Base Paper)

Fig 3 shows that time required for encoding is increases as server count is increase.

TABLE III

Sr. No	Decoding (MS)	Encoding	Number of Scp
1	1	108	4
2	1	116	5
3	2	112	6
4	2	117	7
5	2	115	8

Form above Table II shows encoding & decoding time analysis when share size is K=3. Fig 4 shows that time required for encoding is increases as server count is increase

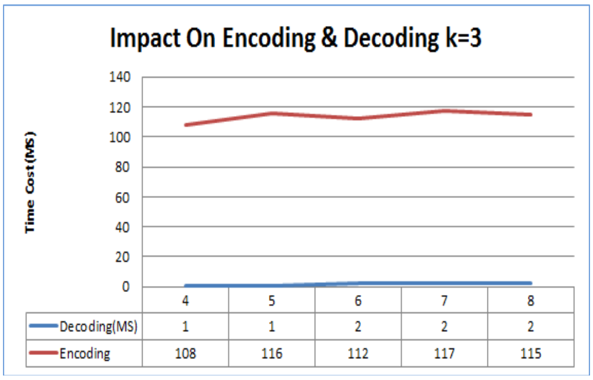

Fig. 4: Impact on Encoding/Decoding time: case 2 (Base Paper)

TABLE IIIII

Sr. No	Encoding(MS)	File Size(KB)
1	36	4
2	43	5
3	51	6
4	63	7
5	71	8
6	83	9

Form above Table III Shows contribution encoding time analysis on different file size. From fig 5 shows that time required for encoding is less than base paper encoding time for same file size. Encoding time is linear to file size.

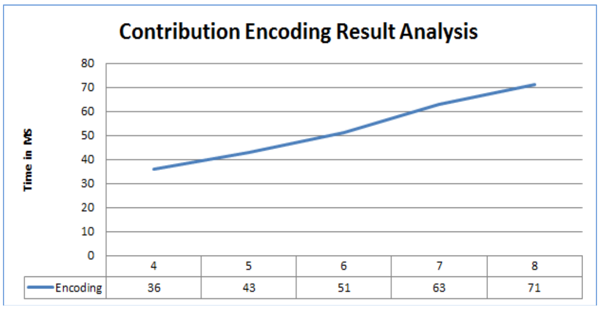

FIG. 5: CONTRIBUTION

From table IV and Figure 6. Show the clear Comparison of the existing system and proposed system contribution. From Result analysis it's clear that Encoding time is less as compared to existing system.

TABLE IVV ENCODING ANALYSIS

Sr. No	Base paper Case1	Base paper Case2	Contribution
1	114	108	36
2	116	116	43
3	118	112	51
4	121	117	63
5	125	115	71
6	125	116	83

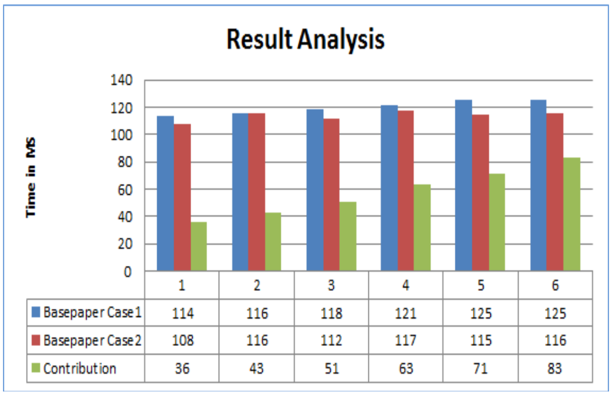

FIG. 6: ENCODING ANALYSIS

VI. CONCLUSION

In the present system a secure and fast sub-graph Similarity Search in Outsourced Cloud Database with data deduplications is evaluated. The prosed system give secure and fast result using attribute based encryption by checking the client credential and sub graph is searched by using apriori algorithm efficiently approach. From the result it's clear that Encoding time is less as compared to existing system.

ACKNOWLEDGMENT

I would like to thanks all people who have helped us directly or indirectly for the paper presentation. I am grateful to my guide, **Prof. P. P. Rokade,** for his guidance and support. I wish to express my sincere thanks to the , PG Coordinator **Prof. I.R. Shaikh** and the departmental staff members for their support. Lastly I would like to thank Department of Computer, SND COE & RC, Yeola for making all the requirements possible and simply available whenever required

REFERENCES

[1] Yun Peng, Zhe Fan, "Authenticated Subgraph Similarity Search in Outsourced Graph Databases,"IEEE Trans. Knowledge and Data Engineering, VOL. 27, NO. 7, JULY 2015.

[2] H. Shang, X. Lin, Y. Zhang, J. X. Yu, and W. Wang, “Connected substructure similarity search,” inProc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 903–914.

[3] Y. Zhu, L. Qin, J. X. Yu, and H. Cheng, “Finding top-k similar graphs in graph databases,” in Proc.15th Int. Conf. Extending Database Technol., 2012, pp. 456–467.

[4] Y. Yuan, G. Wang, L. Chen, and H. Wang, “Efficient subgraph similarity search on large probabilistic graph databases,” in Proc. VLDB Endow., vol. 5, no. 9, pp. 800–811, 2012.

[5] D. W. Williams, J. Huan, and W. Wang, “Graph database indexing using structured graph decomposition,” in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 976–985.

[6] H. He and A. K. Singh, “Closure-tree: An index structure for graph queries,” in Proc. 22nd Int. Conf.ata Eng., 2006, pp. 38–38.

[7] S. Ranu and A. K. Singh, “Indexing and mining topological patterns

[8] for drug discovery,” in Proc. 15th Int. Conf. Extending Database Technol., 2012, pp. 562–565.

[9] D. Shasha, J. Wang, and R. Giugno. Algorithmics and applications of tree and graph searching. In Proc. 21th ACM Symp. on Principles of Database Systems (PODS’02), pages 39–52, 2002.

[10] S. Srinivasa and S. Kumar. A platform based on the multi-dimensional data model for analysis of biomolecular structures. In Proc. 2003 Int. Conf. on Very Large Data Bases, pages 975–986, 2003.

[11] E. Ukkonen. Approximate string matching with q-grams and maximal matches. Theoretic Computer Science, pages 191–211, 1992.

[12] J. Ullmann. Binary n-gram technique for automatic correction of substitution, deletion, insertion, and reversal errors in words. The Computer Journal, 20:141–147, 1977.

[13] J. Wang, K. Zhang, K. Jeong, and D. Shasha. A system for approximate tree matching. IEEE Trans.on Knowledge and Data Engineering, 6:559 – 571, 1994.

[14] 4] P. Willett, J. Barnard, and G. Downs. Chemical similarity searching. J. Chem. Inf. Comput. Sci., 38:983–996, 1998.

[15] Bunke, H., 1997. On a relation between graph edit distance andmaximum common subgraph. Pattern Recognition Lett. 18 _8.,689–694.

[16] Finding Top-K Similar Graphs in Graph Databases Yuanyuan Zhu, Lu Qin, Jeffrey Xu Yu, Hong Cheng

[17] Algorithmic and Applications of Tree and Graph Searching Dennis Shasha Courant Institute New York University

[18] Substructure Similarity Search in Graph Databases. Xifeng Yan IBM T. J. Watson Research Center, psyu@us.ibm.com

[19] F. N. Abu-Khzam, N. F. Samatova, M. A. Rizk, and M. A. Langston, “The maximum common subgraph problem: Faster solu- tions via vertex cover,” in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., 2007, pp. 367–373.

[20] H. Pang, A. Jain, K. Ramamritham, and K.-L. Tan, “Verifying com- pleteness of relational query results in data publishing,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2005, pp. 407–418.

[21] Y. Yang, S. Papadopoulos, D. Papadias, and G. Kollios, “Spatial outsourcing for location-based services,” in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 1082–1091.

[22]] W. Cheng and K.-L. Tan, “Query assurance verification for outsourced multi-dimensional databases,” J. Comput. Secur., vol. 17, no. 1, pp. 101–126, 2009.

[23] Y. Yang, D. Papadias, S. Papadopoulos, and P. Kalnis, “Authenticated join processing in outsourced databases,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 5–18.

[24] F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin, “Dynamic authenticated index structures for outsourced databases,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2006, pp. 121–132.

[25] H. Pang and K. Mouratidis, “Authenticating the query results of text search engines,” Proc. VLDB Endow., vol. 1, no. 1, pp. 126–137, 2008.

[26] M. L. Yiu, Y. Lin, and K. Mouratidis, “Efficient verification of shortest path search via authenticated hints,” in Proc. IEEE 26th Int. Conf. Data Eng., 2010, pp. 237–248.

[27] M. Goodrich, R. Tamassia, N. Triandopoulos, and R. Cohen, “Authenticated data structures for graph and geometric searching,” in Proc. RSA Conf. Cryptographers’ Track, 2003, vol. 2612, pp. 295–313.

[28] A. Kundu and E. Bertino, “How to authenticate graphs without leaking,” in Proc. 13th Int. Conf. Extending Database Technol., 2010, pp. 609–620.

[29] A. Kundu and E. Bertino, “Structural signatures for tree data structures,” in Proc. VLDB Endow., vol. 1, no. 1, pp. 138–150, 2008.

[30] C. Martel, G. Nuckolls, P. Devanbu, M. Gertz, A. Kwong, and S. G. Stubblebine, “A general model for authenticated data structures,” Algorithmica, vol. 39, no. 1, pp. 21–41, Jan. 2004.

[31] H. Jiang, H. Wang, P. Yu, and S. Zhou, “GString: A novel approach for efficient search in graph databases,” in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 566–575.

[32] M. Mann, F. Nahar, H. Ekker, R. Backofen, P. Stadler, and C. Flamm, “Atom mapping with constraint programming,” in Proc. 19th Int. Conf. Principles Practice Constraint Programm., 2013, vol. 8124, pp. 805–822.

[33] Y. Tian and J. M. Patel, “Tale: A tool for approximate large graph matching,” in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 963–972.

[34] A. Khan, N. Li, X. Yan, Z. Guan, S. Chakraborty, and S. Tao, “Neighborhood based fast graph search in large networks,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 901–912.

[35] X. Yan, P. S. Yu, and J. Han, “Graph indexing: A frequent structure-based approach,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2004, pp. 335–346.

BIOGRAPHIES

Nitin D. Dhamale received the B.E. degrees in Computer engineering from the SNJB’s COE, chandwad from Pune University, 2009. Currently Pursuing Master of Degree in Computer engineering from SND COE and RC, Yeola from Pune University.

Prof. P. P. Rokade is presently working as Assistant Professors & Head, Department of Information Technology SNDCOE & RC-Maharashtra. He has presented & attended the papers in several Workshops & Seminars. He has published papers in various National / International Journals.