Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee

Boxiang Dong   Ruilin Liu   Wendy Hui Wang

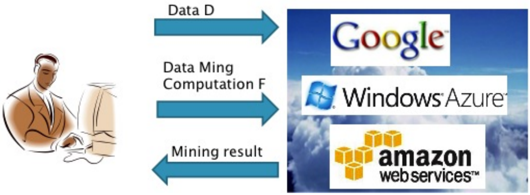
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ

December 10, 2013

STEVENS INSTITUTE *of* TECHNOLOGY
THE INNOVATION UNIVERSITY

Data-mining-as-a-service (DMaS)

Data Mining as a Service:

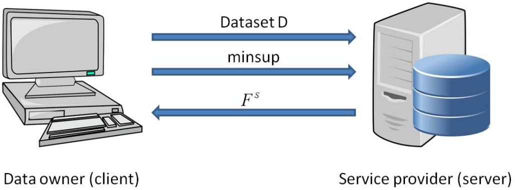


- Weak client
- Computationally powerful service provider (e.g. cloud)
- Result integrity: are the returned mining results the same as if the computation were locally executed?

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Outsourcing Setting

- We focus on the problem of result integrity of outsourced *frequent itemset mining*.
- The architecture of outsourcing frequent itemset mining



Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

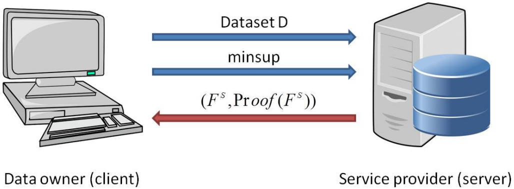
Verification Goal

Given a transaction dataset D and its correct frequent itemset mining result F, let F^S be the errorneous mining result that the server returns.

- Integrity concerns:
  Completeness no frequent itemset is missing in F^S.
  Correctness all itemsets in F^S are frequent.
- We propose an efficient approach to catch incorrect/incomplete mining result with *100% certainty*.

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Verification Framework



- The server constructs cryptographic proofs of the mining results.
    - We use the set intersection verification protocol[PTT11] to construct the proofs.
    - Use the proof to verify the true support of a frequent/infrequent itemset.

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Set Intersection Verification Protocol

Given a collection sets $\mathcal{S} = \{S_1, \ldots, S_m\}$, an intersection result $Y = \{y_1, \ldots, y_\delta\}$, $Y = S_1 \cap S_2 \cap \cdots \cap S_m$ is the *correct* intersection of $\mathcal{S}$ if and only if:

- $(Y \subseteq S_1) \wedge \cdots \wedge (Y \subseteq S_m)$ (subset condition);
- $(S_1 - Y) \cap \cdots \cap (S_m - Y) = \emptyset$ (completeness condition).

Set Intersection Verification Protocol

Given a collection sets $\mathcal{S} = \{S_1, \ldots, S_m\}$, an intersection result $Y = \{y_1, \ldots, y_\delta\}$, $Y = S_1 \cap S_2 \cap \cdots \cap S_m$ is the *correct* intersection of $\mathcal{S}$ if and only if:

- $(Y \subseteq S_1) \wedge \cdots \wedge (Y \subseteq S_m)$ (subset condition);
- $(S_1 - Y) \cap \cdots \cap (S_m - Y) = \emptyset$ (completeness condition).

[PTT11] server prepares $\Pi(Y) = \{\mathcal{B}, \mathcal{A}, \mathcal{W}, \mathcal{C}\}$	client checks
coefficients $\mathcal{B}$ =$\{b_\delta, b_{\delta-1}, \cdots, b_0\}$ of polynomial $(s+y_1)(s+y_2)\cdots(s+y_\delta)$	$\mathcal{B}$ =$\{b_0, \ldots, b_\delta\}$ are correct.
accumulation values $\mathcal{A}$ =$\{acc(S_j) \mid \forall S_j \in \mathcal{S}\}$ where $acc(S_j) = g^{\prod_{x \in S_j}(s+x)}$	$\mathcal{A}$ are correct
subset witness $\mathcal{W} = \{W_j \mid \forall S_j \in \mathcal{S}\}$ where $W_j = g^{P_j(s)}$, $P_j(s) = \prod_{x \in S_j - Y}(x+s)$	$e(\prod_{k=0}^{\lvert Y \rvert}(g^{s^k})^{b_k}, W_j) \stackrel{?}{=} e(acc(S_j), g)$ for $j = 1, \cdots, m$
completeness witness $\mathcal{C} = \{C_j \mid \forall S_j \in \mathcal{S}\}$ for each set $S_j \in \mathcal{S}$, $C_j = g^{q_j(s)}$ s.t. $q_1(s)P_1(s) + q_2(s)P_2(s) + \cdots + q_m(s)P_m(s) = 1$	$\prod_{j=1}^{m} e(W_j, C_j) \stackrel{?}{=} e(g, g)$

Basic Solution

Given a dataset D that contains n unique items, the client does the following:

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Basic Solution

Given a dataset D that contains n unique items, the client does the following:

1. Build the *item-based inverted index* E^I that consists of n inverted lists $\{L_1, \ldots, L_n\}$.
2. Construct the Merkle hash tree $\mathcal{T}$ of the inverted index.
   - Leaf l_j is assigned $h_j = hash(acc(L_j)^{(s+j)})$.
   - Internal node v with children $c_1, \ldots, c_k$ is assigned $h_v = hash(h_{c_1} || \ldots || h_{c_k})$.

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Basic Solution

Given a dataset D that contains n unique items, the client does the following:

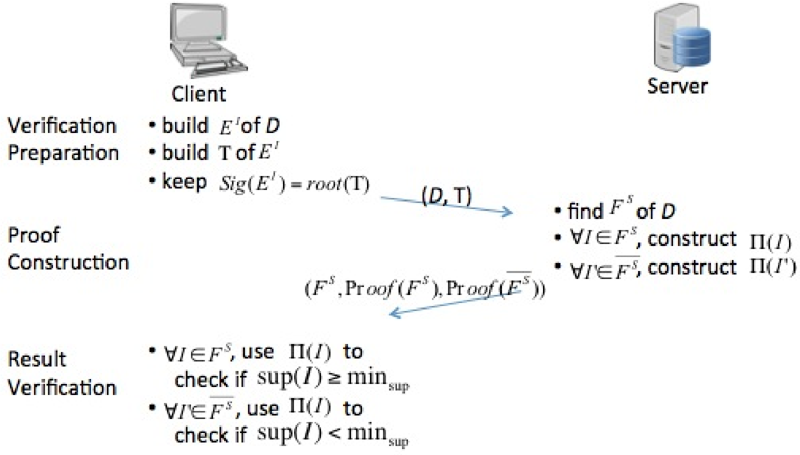
1. Build the *item-based inverted index* E^I that consists of n inverted lists $\{L_1, \ldots, L_n\}$.
2. Construct the Merkle hash tree $\mathcal{T}$ of the inverted index.
   - Leaf l_j is assigned $h_j = hash(acc(L_j)^{(s+j)})$.
   - Internal node v with children $c_1, \ldots, c_k$ is assigned $h_v = hash(h_{c_1}||\ldots||h_{c_k})$.

Mapping to the set intersection verification problem

Verifying whether any itemset I is included in a set of transactions T^I is equivalent to verifying whether T^I is the correct intersection of the inverted lists of all items in I.

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Basic Solution



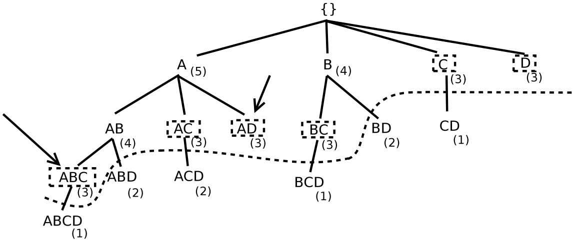
Drawbacks

- Total number of proofs is $2^n - 1$.
- Too much overhead.

Verification Optimization

Maximal frequent itemset (MFI) A subset of F^S s.t. for each itemset $I \in MFI$, there does not exist any itemset $I' \in F^S$ s.t. $I \subseteq I'$.

Minimal infrequent itemset (MII) A set of itemsets that do not appear in F^S s.t. for each itemset $I \in MII$, there does not exist any itemset $I' \notin F^S$ s.t. $I' \subseteq I$.



(Itemsets in dotted rectangles are maximal frequent itemsets.)

Advantage $|MFI| + |MII| \ll |F^S| + |\overline{F^S}|$

Optimized Solution

Client

Server

Verification Preparation

- build E^I of *D*
- build T of E^I
- keep $Sig(E^I) = root(\mathrm{T})$

Client → Server: (*D*, T)

Proof Construction

- find F^S of *D*
- find *MFI* and *MII*
- $\forall I \in MFI$, construct $\Pi(I)$
- $\forall I' \in MII$, construct $\Pi(I')$

Server → Client: $(F^S, \mathrm{Pr}\,oof(MFI), \mathrm{Pr}\,oof(MII))$

Result Verification

- correctness verification with *MFI*
- completeness verification with *MII*

Security Analysis Our optimized solution provides the same security guarantee as the basic solution.

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Complexity

Proof construction at server side $O(Mlog^3 M + n^\epsilon logn)$

- $M = \sum_{I \in MFI \cup MII} \sum_{i \in I} |L_i|$
- n is the number of unique items of D.
- $\epsilon \in (0, 1)$

Verification at client side $O(N + F)$

- $N = \sum_{I \in MFI \cup MII} |I|$
- $F = \sum_{I \in MFI \cup MII} sup(I)$

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Experiments

- Environment

  Language C++

  Testbed Macbook Pro, 2.4GHz CPU, 4 GB memory

-

Dataset	# of trans.	# of items	Avg. trans. length	min_{sup}	# of freq. itemsets
S_1	10^3	49	10	250	36
S_2	10^4	49	10	250	3854
S_3	10^5	49	10	250	149744
S_4	10^6	49	10	250	3074610
R	500	100	2.4	5	97

- Simulation of malicious actions

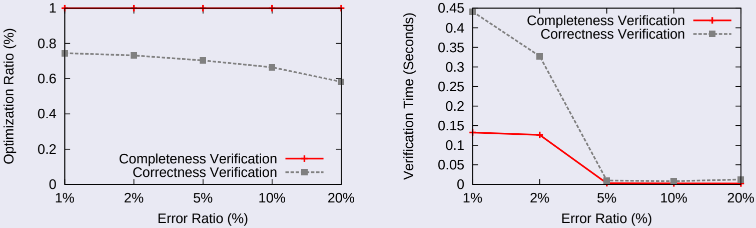
  Error ratio $r = 1\%, 2\%, 5\%, 10\%, 20\%$

  Incomplete Randomly delete r percent mining result.

  Incorrect Randomly insert r percent infrequent itemsets.

Proof Optimization Ratio & Verification Time

Optimization Ratio & Verification Time (*R* dataset)



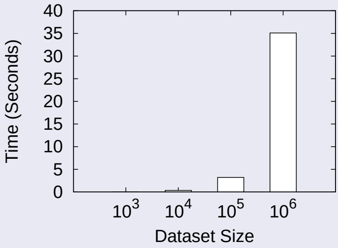
(a) Proof optimization ratio

(b) Client verification time

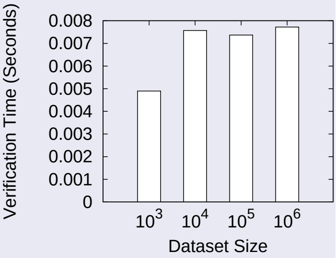
Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Scalability

Scalability (error ratio=1%)



(a) Construction time of one proof (itemset length = 3)



(b) Client verification time

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

References I

[Bab85] László Babai.
Trading group theory for randomness.
In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 421–429. ACM, 1985.

[DLW13] Boxiang Dong, Ruilin Liu, and Hui Wendy Wang.
Result integrity verification of outsourced frequent itemset mining.
In *Data and Applications Security and Privacy XXVII*, pages 258–265. Springer, 2013.

[GGP10] Rosario Gennaro, Craig Gentry, and Bryan Parno.
Non-interactive verifiable computing: Outsourcing computation to untrusted workers.
In *Advances in Cryptology–CRYPTO 2010*, pages 465–482. Springer, 2010.

[GMR89] Shafi Goldwasser, Silvio Micali, and Charles Rackoff.
The knowledge complexity of interactive proof systems.
SIAM Journal on computing, 18(1):186–208, 1989.

[LWM+12] Ruilin Liu, Hui Wendy Wang, Anna Monreale, Dino Pedreschi, Fosca Giannotti, and Wenge Guo.
Audio: an integrity auditing framework of outlier-mining-as-a-service systems.
In *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases-Volume Part II*, pages 1–18. Springer-Verlag, 2012.

[PJRT05] HweeHwa Pang, Arpit Jain, Krithi Ramamritham, and Kian-Lee Tan.
Verifying completeness of relational query results in data publishing.
In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 407–418. ACM, 2005.

References II

[PRV12] Bryan Parno, Mariana Raykova, and Vinod Vaikuntanathan.
How to delegate and verify in public: Verifiable computation from attribute-based encryption.
In *Theory of Cryptography*, pages 422–439. Springer, 2012.

[PTT11] Charalampos Papamanthou, Roberto Tamassia, and Nikos Triandopoulos.
Optimal verification of operations on dynamic sets.
In *Advances in Cryptology–CRYPTO 2011*, pages 91–110. Springer, 2011.

[RHPH13] Liu Ruilin, (Wendy) Wang Hui, Mordohai Philippos, and Xiong Hui.
Integrity verification of k-means clustering outsourced to infrastructure as a service (iaas) providers.
In *Proceedings of 2013 SIAM International Conference on Data Mining (SDM)*, pages 632–640. SIAM, 2013.

[Sio05] Radu Sion.
Query execution assurance for outsourced databases.
In *Proceedings of the 31st international conference on Very large data bases*, pages 601–612. VLDB Endowment, 2005.

[WCH+09] Wai Kit Wong, David W Cheung, Edward Hung, Ben Kao, and Nikos Mamoulis.
An audit environment for outsourcing of frequent itemset mining.
Proceedings of the VLDB Endowment, 2(1):1162–1173, 2009.

[XWYM07] Min Xie, Haixun Wang, Jian Yin, and Xiaofeng Meng.
Integrity auditing of outsourced data.
In *Proceedings of the 33rd international conference on Very large data bases*, pages 782–793. VLDB Endowment, 2007.

Q & A

Thank you!

Questions?

Related Work

Verifiable Computation

- [Bab85, GMR89, PRV12, GGP10] the expensive pre-processing phase is amortized over the future executions.

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Related Work

Verifiable Computation

- [Bab85, GMR89, PRV12, GGP10] the expensive pre-processing phase is amortized over the future executions.

Integrity Verification of Database-as-a-Service (DaS)

- [PJRT05, Sio05, XWYM07] provide assurance for SQL query results.

Related Work

Verifiable Computation

- [Bab85, GMR89, PRV12, GGP10] the expensive pre-processing phase is amortized over the future executions.

Integrity Verification of Database-as-a-Service (DaS)

- [PJRT05, Sio05, XWYM07] provide assurance for SQL query results.

Integrity Verification of DMaS

- [WCH+09, DLW13] only provide probabilistic result integrity guarantee.
- [LWM+12, RHPH13] focus on other mining tasks (outlier detection, clustering)

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang

Client versus Server

Comparison on S_1 dataset

min_{sup}	# of Freq. Itemsets	Client side	Server side	
		Verify	Proof prep.	mining
402	10	0.000164	24.72	0.03707
203	50	0.001358	266.985	0.08984
157	99	0.00332	572.591	0.1355

(time measured in seconds)

Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee. ICDM 13. Dong, Liu, Wang