T-79.515 Special Course on Cryptology

Seminar I: Secure Frequent Itemset Mining

Sven Laur

Helsinki University of Technology

swen@math.ut.ee,slaur@tcs.hut.fi

Overview

- The problem and motivation
- Briefly about (distributed) Apriori
- Information leaks of distributed Apriori
- Private union protocol
- Private addition—Benaloh' protocol
- Remarks about two-party setting

What are frequent sets and association rules?

Database $\mathcal{DB}$ is a list of records R.
Each record $R = \{\mathcal{I}_1, \ldots, \mathcal{I}_k\}$ is a subset of items I.

- The support of the itemset $\mathcal{A}$ is

$$\mathsf{supp}(\mathcal{A}) = \#\{R \in \mathcal{DB} \,:\, \mathcal{A} \subseteq R\}\,.$$

- The support of the association rule $\mathcal{A} \Rightarrow \mathcal{B}$

$$\mathsf{supp}(\mathcal{A} \Rightarrow \mathcal{B}) = \mathsf{supp}(\mathcal{A} \cup \mathcal{B}).$$

- The confidence of the association rule $\mathcal{A} \Rightarrow \mathcal{B}$

$$\mathsf{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\mathsf{supp}(\mathcal{A} \cup \mathcal{B})}{\mathsf{supp}(\mathcal{A})}.$$

Frequent itemset mining is sufficient

- Only the rules with sufficient support are interesting.

- Frequent sets reveal all frequent association rules.

Cooperative frequent set mining

Usually large data collection is divided:

- horizontally — records are not divided;
- vertically — different parties have parts of record.

Consider horizontally partitioned database $\mathcal{DB} = \mathcal{DB}_1 \cup \mathcal{DB}_2 \cup \ldots \cup \mathcal{DB}_t$.

- Only local association rules are available to each party.
- Parties must share information to find global association rules.
- Parties do not trust each other.
- How to find global rules without revealing local data and meta-data.

Data mining with an independent referee

A well-established independent referee does data mining.
The referee is *conditionally trusted party*

- We trust that computed results are correct and published.
- The referee may sell intermediate results to other parties.
- Parties do not trust each other.
- How to find global rules without revealing local data and meta-data.
- Can we device more efficient protocol.

How to mine frequent sets?

The key ingredient of the Apriori algorithm is anti-monotone relation

$$\mathcal{A} \subseteq \mathcal{B} \qquad \Longrightarrow \qquad \mathsf{supp}(\mathcal{A}) \geq \mathsf{supp}(\mathcal{B}).$$

Subsets of a frequent set are also frequent sets!

Principle of Apriori:

- find frequent one-element itemsets;
- find frequent two-element itemsets;
- ...
- no candidates for ℓ-element itemsets, halt.

Apriori in a pure form

Input: Support threshold κ.

$\spadesuit\spadesuit\spadesuit$ No frequent sets, candidate set I $\spadesuit\spadesuit\spadesuit$

$F = \emptyset$; $C = I$; $\ell = 1$

while $|C| > 0$ **do**

$\spadesuit\spadesuit\spadesuit$ Find all valid candidates $\spadesuit\spadesuit\spadesuit$

$F = F \cup \{\mathcal{A} \in C : \ \mathsf{supp}(\mathcal{A}) > \kappa\}$

$\spadesuit\spadesuit\spadesuit$ Form ℓ-element candidate set $\spadesuit\spadesuit\spadesuit$

$C = \{\mathcal{B} \in \mathcal{P}(I) : \ |\mathcal{B}| = \ell, \ \mathcal{A} \subseteq \mathcal{B} \Rightarrow \mathcal{A} \in F\}$

$\ell = \ell + 1$

return F

How to mine frequent sets in distributed setting?

Let $n = |\mathcal{DB}|$ and $n_i = |\mathcal{DB}_i|$. Then the following implication holds

$$\mathsf{supp}(\mathcal{B}) > \kappa \qquad \Longrightarrow \qquad \exists i :\ \mathsf{supp}_i(\mathcal{B}) > \frac{n_i \kappa}{n} = \kappa_i .$$

Three classes of frequent itemsets:

$$F = \{\mathcal{A} :\ \mathsf{supp}(\mathcal{A}) > \kappa\}, \qquad F_i = \{\mathcal{A} :\ \mathsf{supp}_i(\mathcal{A}) > \kappa_i\},$$
$$LF_i = F \cap F_i = \{\mathcal{A} :\ \mathsf{supp}(\mathcal{A}) > \kappa,\ \mathsf{supp}_i(\mathcal{A}) > \kappa_i\}.$$

If $\mathcal{B}$ globally frequent itemset, then following holds

$$\mathcal{B} \in F \qquad \Longrightarrow \qquad \exists i :\ \mathcal{A} \subseteq \mathcal{B} \Rightarrow A \in LF_i .$$

We can deal only with locally supported globally frequent sets!

Distributed Apriori in a pure form

Input: Normalized support threshold κ/n.

♠♠♠ Calculate local threshold ♠♠♠
$\kappa_i = n_i\kappa/n$
♠♠♠ No frequent sets, candidate set I ♠♠♠
$F = \emptyset$; $LF_i = \emptyset$; $C_i = I$; $C = I$; $\ell = 1$
while $|C| > 0$ **do**

  ♠♠♠ Find all valid candidates ♠♠♠
  $F_i^* = \{\mathcal{A} \in C_i : \mathsf{supp}_i(\mathcal{A}) > \kappa_i\}$
  ♠♠♠ Broadcast candidates ♠♠♠
  $C = F_1^* \cup \cdots \cup F_t^*$
  ♠♠♠ Global test ♠♠♠
  $F = F \cup \{\mathcal{B} \in C : \mathsf{supp}(\mathcal{B}) > \kappa\}$
  $LF_i = LF_i \cup \{\mathcal{B} \in C : \mathsf{supp}(\mathcal{B}) > \kappa, \mathsf{supp}_i(\mathcal{B}) > \kappa_i\}$
  ♠♠♠ New local candidate set ♠♠♠
  $C_i = \{\mathcal{B} \in \mathcal{P}(I) : |\mathcal{B}| = \ell, \mathcal{A} \subseteq \mathcal{B} \Rightarrow \mathcal{A} \in LF_i\}$
  $\ell = \ell + 1$

return F

Private union protocol (Clifton and Kantarcioglu)

The protocol is based on a commutative encryption scheme that is

$$\mathsf{E}_1\mathsf{E}_2\ldots\mathsf{E}_t(\mathcal{A}) = \mathsf{E}_{\pi(1)}\mathsf{E}_{\pi(2)}\ldots\mathsf{E}_{\pi(t)}(\mathcal{A})$$

for all possible messages and permutations π. The probability of collisions

$$\Pr\left[\mathsf{E}_1\mathsf{E}_2\ldots\mathsf{E}_t(\mathcal{A}_1) = \mathsf{E}_{\pi(1)}\mathsf{E}_{\pi(2)}\ldots\mathsf{E}_{\pi(t)}(\mathcal{A}_2)\right],$$

when $\mathcal{A}_1 \neq \mathcal{A}_2$, should be negligible.

Given

$$\mathsf{E}_1\ldots\mathsf{E}_t(\{\mathcal{A}_1,\ldots,\mathcal{A}_k\}) \quad \text{and} \quad \mathsf{E}_{\pi(1)}\ldots\mathsf{E}_{\pi(t)}(\{\mathcal{B}_1,\ldots,\mathcal{B}_k\}),$$

we can eliminate duplicates $\mathcal{A}_i = \mathcal{B}_j$.

Security of the CK protocol

The CK protocol privately computes the union if there are no colluding parties and reveals at most:

- size of all intersections $|C_i \cap C_{i+2k}|$;
- size of intersection $|D_1 \cap D_2|$;
- size of $|D_1|$ and $|D_2|$;

where $D_1 = C_1 \cup C_3 \cup \ldots$ and $D_2 = C_2 \cup C_4 \cup \ldots$.

Re-execution allows parties 1 and 2 to distinguish repeating sets.

Generic union protocol

Public input: Superset X.

Private input: Set $C_i = \left\{\mathcal{A}_1, \cdots, \mathcal{A}_{k_i}\right\}$.

$C = \emptyset$

for $\mathcal{A} \in X$ **do**

♠♠♠ $\mathcal{A} \in C_1 \vee \ldots \vee \mathcal{A} \in C_t = \neg(\neg(\mathcal{A} \in C_1) \wedge \ldots \wedge \neg(\mathcal{A} \in C_t))$ ♠♠♠

if $\mathcal{A} \in C_i$ **then** $b_i = 0$ **else** $b_i = 1$

Securely multiply $c \equiv b_1 \cdots b_t \mod 2$.

if $c \neq 1$ **then**

Add $\mathcal{A}$ to C.

return C

Benaloh' protocol

Random matrix in additive (multiplicative) group G

$$
\begin{array}{c|c}
a_1 & \\
a_2 & \\
\vdots & \\
a_t &
\end{array}
\begin{array}{c}
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1t} \\
a_{21} & a_{22} & \cdots & a_{2t} \\
\vdots & \vdots & \ddots & \vdots \\
a_{t1} & a_{t2} & \cdots & a_{tt}
\end{pmatrix} \\
\hline
\begin{matrix} b_1 & b_2 & \dots & b_t \end{matrix}
\end{array}
$$

- Row sums a_i are fixed.
- All proper subset of row elements have uniform distribution.
- Column sums of $t-1$ arbitrary columns have also uniform distribution.
- Sum of column sums is $a = b_1 + \cdots + b_t$.

Benaloh' protocol in a pure form

Private input: Private term a_i.

Choose randomly $a_{i1} + a_{i2} + \cdots + a_{it} = a_i$.
for $j = 1$ ***to*** *t* **do**
  Send a_{ij} to the jth party.

Calculate column sums $b_i = a_{1i} + a_{2i} \cdots + a_{ti}$
Broadcast values b_i.

return $b_1 + b_2 + \cdots + b_t$

Security of the Benaloh' protocol

The Beneloh' protocol is unconditionally secure against coalition up to $t-1$ parties.

The generic union protocol that uses Benaloh' protocol for multiplication is unconditionally secure against coalition up to $t-1$ parties.

- The generic union protocol is computationally more efficient.
- The generic union protocol has large communication complexity.
- The security quarantee is more strict!

Secure threshold test

We need to evaluate predicate $\mathsf{supp}_1(\mathcal{A}) + \cdots + \mathsf{supp}_1(\mathcal{A}) > \kappa$?.
Naive solution assuming that there are no coalitions.

- Parties 1 and t are special.
- Party 1 starts summing procedure $s = \mathsf{supp}_1(\mathcal{A}) - \kappa_1 + r$.
- Other parties add their shares $s = s + \mathsf{supp}_i(\mathcal{A}) - \kappa_i$.
- Parties 1 and t test whether $s - r > 0$? with Yao's circuit.

Coalitions break the protocol down!

CK inequality test

Private input: Private support $\mathsf{supp}_i(\mathcal{A})$.

Public input: Large modulus $m > 2n$ such that $\mathsf{gcd}(m, n) = 1$.

Party 1 chooses $r \in \mathbb{Z}_m$.
Sets $c \equiv \mathsf{supp}_1(\mathcal{A}) + r - \kappa_1 \mod m$.
for $i = 1$ ***to*** t **do**
$\quad c \equiv c + \mathsf{supp}_i(\mathcal{A}) - \kappa_i \mod m$.

Parties 1 and t use Yao's circuit and determine ?$c - r \geq 0 \mod m$.

The condition $\mathsf{gcd}(m, n) = 1$ allows to embed fractional thresholds κ_i

$$\kappa_1 + \kappa_2 + \cdots + \kappa_t \equiv \kappa \mod m.$$

Two-party case is futile

- A global support reveals the local support of the other party.
- If the hostile party provides empty database or uniformly filled database, the he can deduce all frequent sets of the victim.
- There are no feasible cryptographic mechanisms to prevent the attack!