



**Purdue University**  
**Center for Education and Research in**  
**Information Assurance and Security**



---

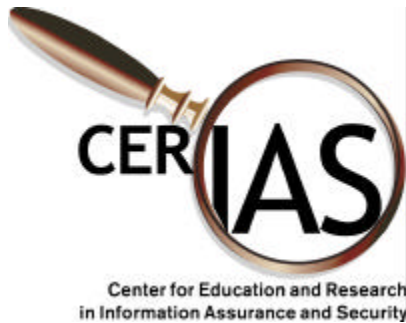
# Association Rule Hiding

---

Elena Dasseni and Yucel Saygin

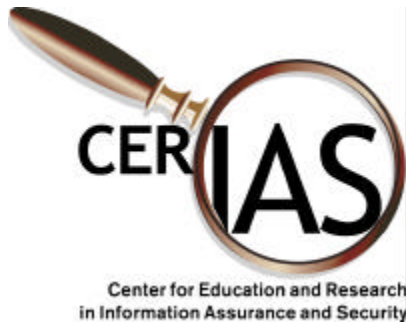
Contributors:

M. Atallah, E. Bertino, A. Elmagarmid, V. Verykios M. Ibrahim



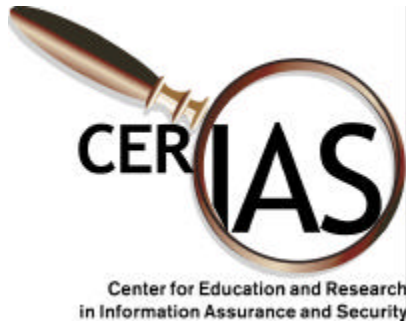
# Introduction

- Restricting access to sensitive data and the “inference” problem.
- Security risks due to recent advances in data mining techniques.
- Association Rules (i.e., “90% of air-force basis having super-secret plane A, also have helicopters of type B”).



## Introduction(Contd.)

- Security and privacy threats from data mining and similar applications.
- Possible solutions to prevent data mining of significant knowledge:
  - Releasing only subsets of the source database
  - Augmenting the database
  - Disclosing an aggregated but not individual value



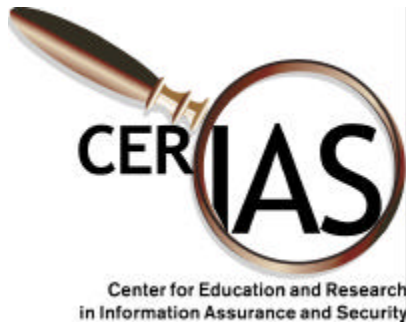
# Association Rule Discovery

Let  $I = \{i_1, i_2, \dots, i_m\}$  be a set of literals, called items.

A set of items  $X \subset I$  is called an itemset.

Let  $D$  be a set of transactions, where each transaction  $T$  is an itemset such that  $T \subseteq I$ .

A transaction  $T$  contains an itemset  $X$ , if  $X \subseteq T$ .



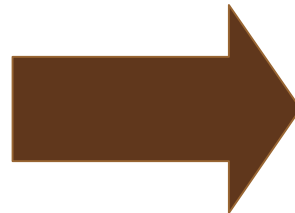
# Association Rule Discovery

An association rule is an implication of the form  $X \Rightarrow Y$  where  $X \subset I$ ,  $Y \subset I$ , and  $X \cap Y = \emptyset$ .

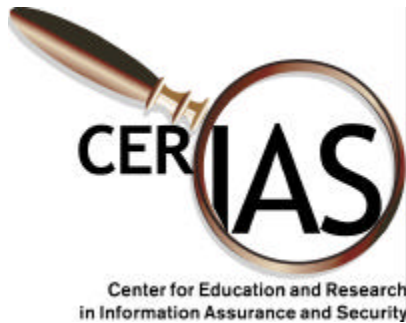
$$\text{confidence} = \frac{|X \cup Y|}{|X|}, \text{ and support} = \frac{|X \cup Y|}{N}$$

Example Database

TID	Items
T1	ABCD
T2	ABC
T3	ACD

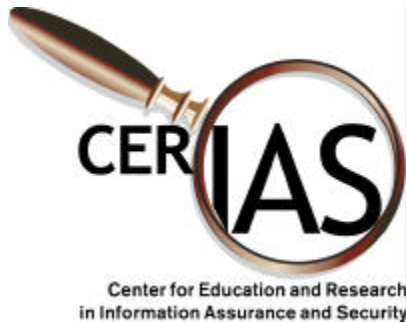


Frequent Itemsets	Support
AB	2
AC	3
AD	2
BC	2
CD	2
ABC	2
ACD	2



# Optimal Sanitization is NP-hard

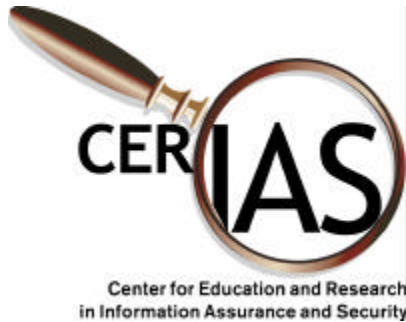
- Let  $D$  be the source database.
- Let  $R$  be a set of “significant” association rules that are mined from  $D$ .
- Let  $r_i$  be a “sensitive” rule in  $R$ .
- Transform  $D$  into  $D'$  so that all rules in  $R$  can still be mined from  $D'$  but  $r_i$ .
- Optimal sanitization is NP-Hard.
- Reduction from the NP-Hard problem of Hitting Set.



## Hiding Methods

- Reduce the support of frequent itemsets containing sensitive rules
  - Cyclic Method
  - Greedy Method
  - Isolated items and safe transactions
- Reduce the confidence or support of rules

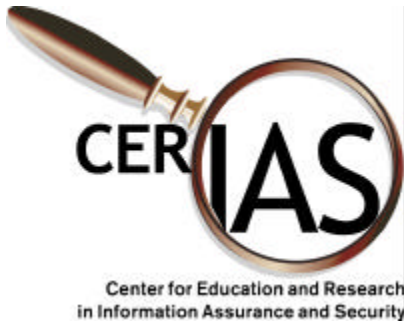




## Hiding Association Rules by using Confidence and Support

- **Assumptions**

- We hide a rule by decreasing either its confidence or its support
- We decrease either the support or the confidence one unit at a time (we modify the value of one transaction at a time)
- We hide one rule at a time
- We consider only set of disjoint rules (rules supported by large itemsets that do not have any common item)



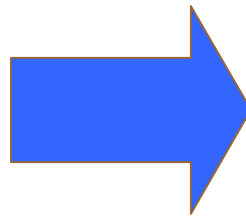
## Hiding a rule $X \rightarrow Y$ by using Confidence and Support

- **$\text{Conf}(X \rightarrow Y) = \text{Supp}(XY) / \text{Supp}(X)$**
- **Strategies**
  - Decreasing confidence of rule
    - Increasing the support of  $X$  in transactions not supporting  $Y$
    - Decreasing the support of  $Y$  in transactions supporting both  $X$  and  $Y$
  - Decreasing support of rule
    - Decreasing the support of the corresponding large itemset ( $XY$ )

## Strategies: basic idea

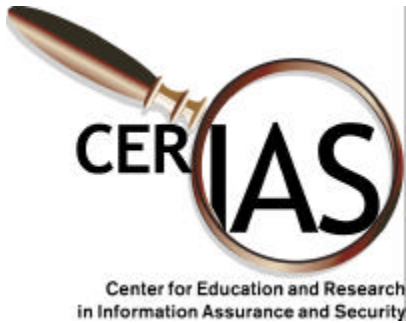
- Transactions viewed as lists
- One element for each item in DB

TID	Items
T1	ABC
T2	A



TID	A	B	C
T1	1	1	1
T2	1	0	0

- Decreasing support of  $S$  = turning to 0 one item in one transaction supporting  $S$
- Increasing support of  $S$  = turning to 1 one item in one transaction partially supporting  $S$



## Example

$$\text{MIN\_SUPP} = 1/5 = 20\%$$

$$\text{MIN\_CONF} = 80\%$$

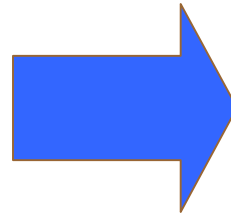
TID	Items
T1	ABC
T2	ABC
T3	A C
T4	A
T5	B

AR	Conf
AB→C	100%
BC→A	100%

## Example: hiding $AB \rightarrow C$ by increasing support of AB

- Turn to 1 the item B in transaction T4

TID	Items
T1	ABC
T2	ABC
T3	A C
T4	A
T5	B



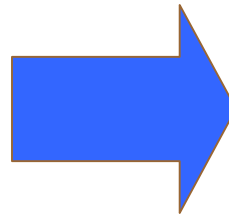
TID	Items
T1	ABC
T2	ABC
T3	A C
T4	AB
T5	B

AR	Conf
$AB \rightarrow C$	66%
$BC \rightarrow A$	100%

## Example: hiding $AB \rightarrow C$ by decreasing support of C

- Turn to 0 the item C in T1

TID	Items
T1	ABC
T2	ABC
T3	A C
T4	A
T5	B



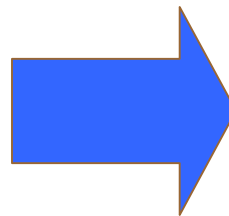
TID	Items
T1	AB
T2	ABC
T3	A C
T4	A
T5	B

AR	Conf
$AB \rightarrow C$	50%
$BC \rightarrow A$	100%

## Example: hiding $AB \rightarrow C$ by decreasing support of ABC

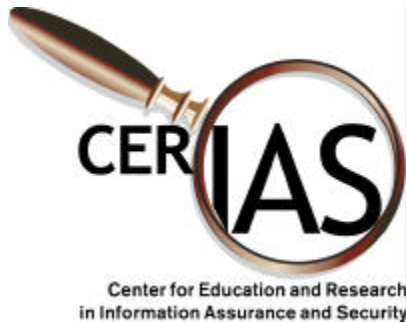
- Turn to 0 the item B in T1
- Turn to 0 the item C in T2

TID	Items
T1	ABC
T2	ABC
T3	A C
T4	A
T5	B



TID	Items
T1	A C
T2	AB
T3	A C
T4	A
T5	B

AR	Conf
$AB \rightarrow C$	0%
$BC \rightarrow A$	0%



## Conclusions

- DM as a threat to DB security
- Need to limit the disclosure of sensitive information
- Optimal sanitization is NP-hard
- Developed heuristics to solve the problem
- The proposed methods are implemented and tested
- We plan to extend the problem of limiting the disclosure of sensitive information for different data mining techniques