CERIAS

IS DATA MINING DANGEROUS?

A study on the data mining effects on the anonymity of individuals

Maurizio Atzori[1,2], Francesco Bonchi[2], Fosca Giannotti[2], Dino Pedreschi[3]

[1]Purdue University *(Visiting Scholar)*, [2]ISTI-CNR Pisa, [3]University of Pisa

Data mining

Data Mining (*DM*) is the process of extracting useful information (e.g., *rules*) from large amount of data.

Data anonymity

Data are considered anonymous if you cannot link them to people. ID and quasi-ID (subset of public available attributes that can be used as ID) need to be masked or removed.

Can data mining results violate anonymity of individuals?

Surprisingly **yes, they can**!

Individually secure

Age = 27, Postcode = 45254, Christian ⇒ American
(support = 758, confidence = 99.8%)

Age = 27, Postcode = 45254 ⇒ American
(support = 1053, confidence = 99.9%)

Since *sup(rule) / conf(rule) = sup(head)* we can derive:

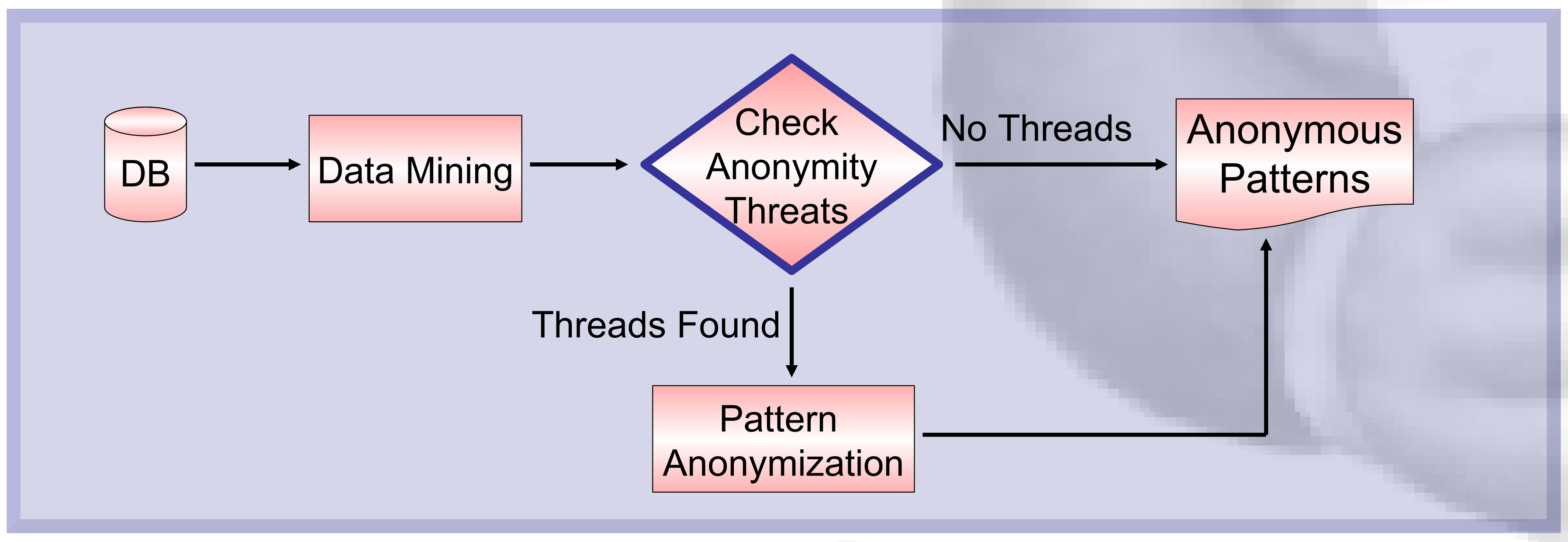
Dangerous!

Age = 27, Postcode = 45254, not American ⇒ Christian
(support = 1, confidence = 100.0%)

This information refers to my German neighbor, he is Christian! (and this information was clearly not intended to be released as it links public information regarding few people to sensitive data!)

How to find all anonymity breaches in DM results?

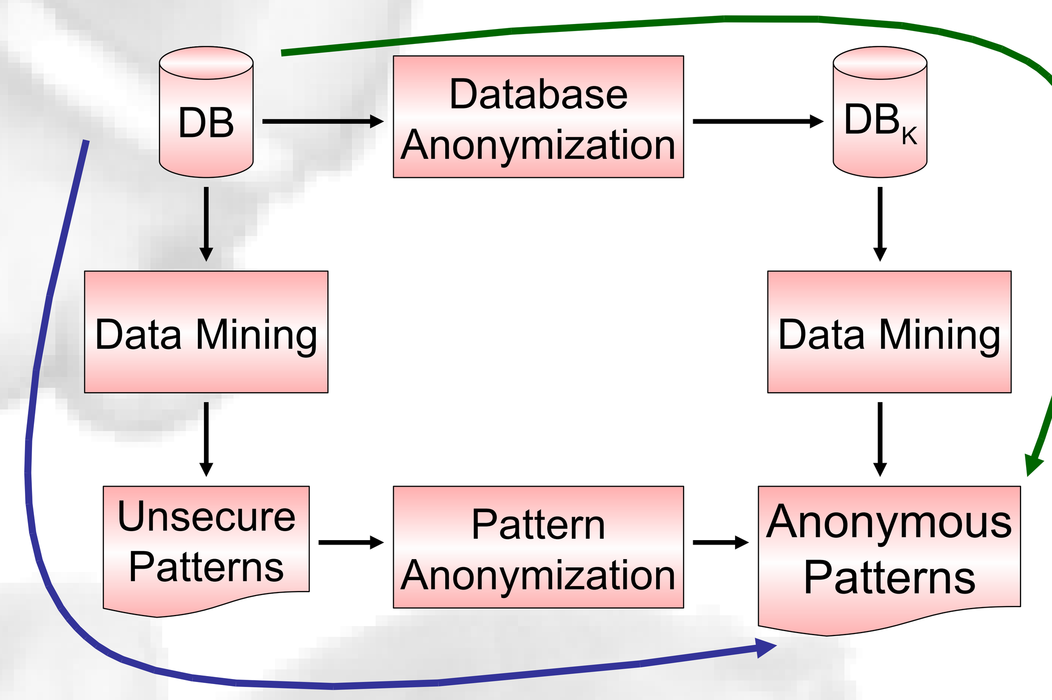
We developed naïve and optimized ***inference channel detectors*** that exploit theoretical results on closed sets.



How to get anonymity-preserving DM results?

Two alternatives:

1. Anonymize the DB, then do mining
2. **Do usual mining, then *block anonymity threats***



… and the second path preserves more information

Enforcing Pattern Anonymity

ADD and *SUP* algorithms can be used to block anonymity threats, by merging inference channels and then modifying the original support of patterns.

ADD increments the support of infrequent patterns, while SUP suppresses the information about infrequent data.

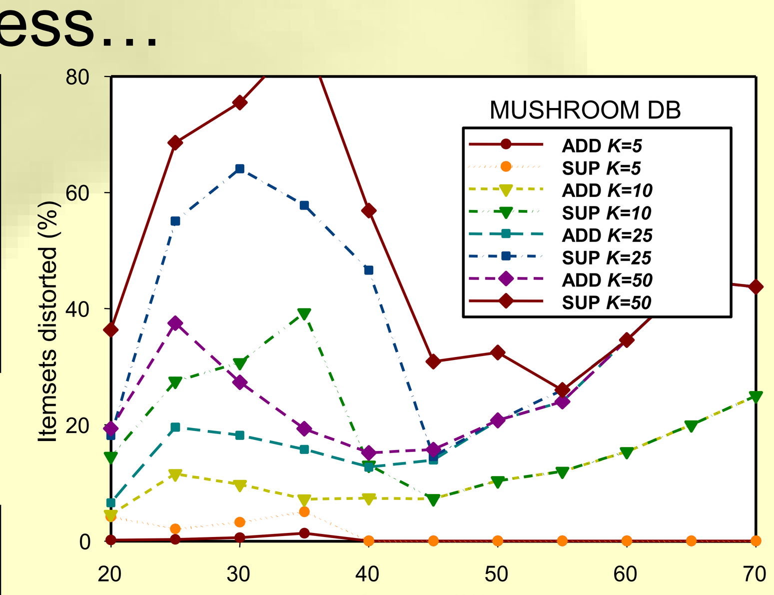
Both are shown to preserve information while removing anonymity threats.

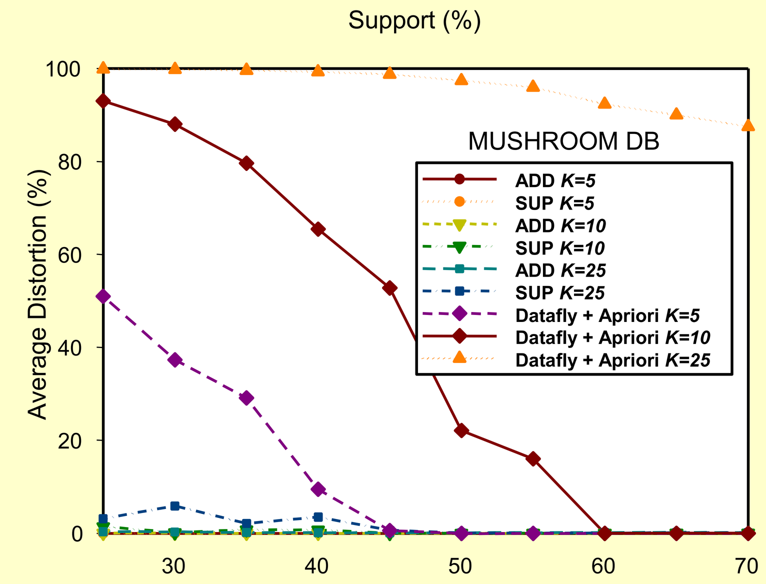
Mining results from anonymized database can be useless…

Pattern	Support
Age = Any, Native-Country = Any, Race = Any, Hours-per-Week = 2	18577
Age = Any, Native-Country = Any, Race = Any, Sex = Male, Capital-Gain = 0	18403
Age = Any, Native-Country = Any, Race = Any, Sex = Male, Capital-Loss = 0	19290
Age = Any, Native-Country = Any, Race = Any, WorkClass = Private, Capital-Gain = 0, Capital-Loss = 0	19623
Age = Any, Native-Country = Any, Race = Any, Income = Low, Capital-Gain = 0, Capital-Loss = 0	21021

…while pattern anonymization preserves information!

Pattern	Support	ADD	SUP
Native-Country = United-States, Capital-Loss = 0, WorkClass = Private	19237	19237	19237
Capital-Loss = 0, Sex = Male	19290	19290	19290
Race = White, Capital-Gain = 0, Income = Low	18275	18275	18249
Race = White, Capital-Loss = 0, Income = Low	18489	18489	18489
Sex = Male, Capital-Gain = 0	18403	18403	18263
Race = White, Capital-Loss = 0, WorkClass = Private	18273	18273	18273
Native-Country = United-States, Sex = Male	18572	18572	18512
Race = White, Native-Country = United-States, Capital-Loss = 0, Capital-Gain = 0	20836	20836	20836
Native-Country = United-States, WorkClass = Private, Capital-Gain = 0	18558	18558	18493
Hours-per-Week = 2	18557	18557	18446
Capital-Loss = 0, WorkClass = Private, Capital-Gain = 0	19623	19623	19573
Native-Country = United-States, Capital-Loss = 0, Capital-Gain = 0, Income = Low	19009	19009	19009
Number of tuples in the database (Adult DB)	***32162***	***30212***	***29967***





1. Atzori et al., *K-Anonymous Patterns*, PKDD05
2. Atzori et al., *Blocking Anonymity Threats Raised by Frequent Itemset Mining*, ICDM05

Pisa KDD Laboratory - http://www-kdd.isti.cnr.it/

PURDUE UNIVERSITY · PURDUE HOMELAND SECURITY INSTITUTE · CERIAS · REGENSTRIEF CENTER FOR HEALTHCARE ENGINEERING · Discovery Park e-Enterprise Center