# *k*-Anonymous Patterns

M. Atzori[1,2] F. Bonchi[2] F. Giannotti[2] D. Pedreschi[1]

[1]Department of Computer Science
University of P*i*sa

[2]Information Science and Technology Institute
CNR, P*i*sa

*Speaker: Maurizio Atzori*

9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2005

# Outline

# Outline

# Outline

# A Taxonomy of Privacy Preserving Data Mining

1. Intesional Knowledge Hiding
   - Bertino's approach, based on DB Sanitization
2. Extensional Knowledge Hiding
   - Agrawal's approach, based on DB Randomization
   - Sweeney's approach, based on DB Anonymization
3. Distributed Extensional Knowledge Hiding
   - Clifton's approach based on Secure Multiparty Computation
4. Secure Intesional Knowledge Sharing
   - Clifton's Public/Private/Unknown Attribute Framework
   - Zaïane's Association Rule Sanitization (but also IKH)
   - *k*-Anonymous Patterns – Our approach

# The Purpose

- We want to publish datamining results (like Secure Intesional Knowledge Sharing)
- We DON'T want to release information related to few people, that can help to trace single individuals
- We don't want to specify any other information

# A Motivating Example in the Medical Domain

## Example

- Suppose Dr. Gregory House conduces both usual hospital activities and research

# A Motivating Example in the Medical Domain

## Example

- Suppose Dr. Gregory House conduces both usual hospital activities and research
- He has a big database with all sensitive information about his patients

# A Motivating Example in the Medical Domain

## Example

- Suppose Dr. Gregory House conduces both usual hospital activities and research
- He has a big database with all sensitive information about his patients
- Playing with Data Mining, he discovered interesting trends about patologies in his patient data

# A Motivating Example in the Medical Domain

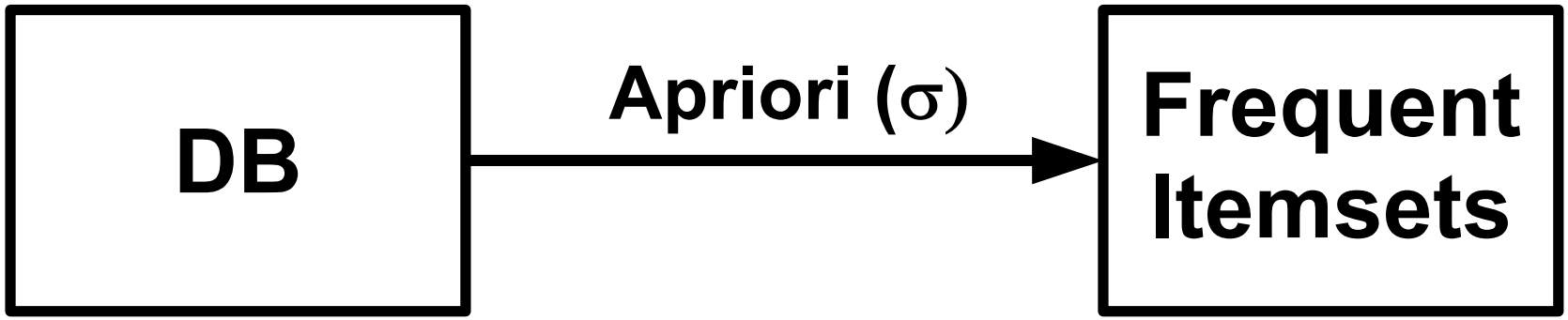
## Example

- Suppose Dr. Gregory House conduces both usual hospital activities and research
- He has a big database with all sensitive information about his patients
- Playing with Data Mining, he discovered interesting trends about patologies in his patient data

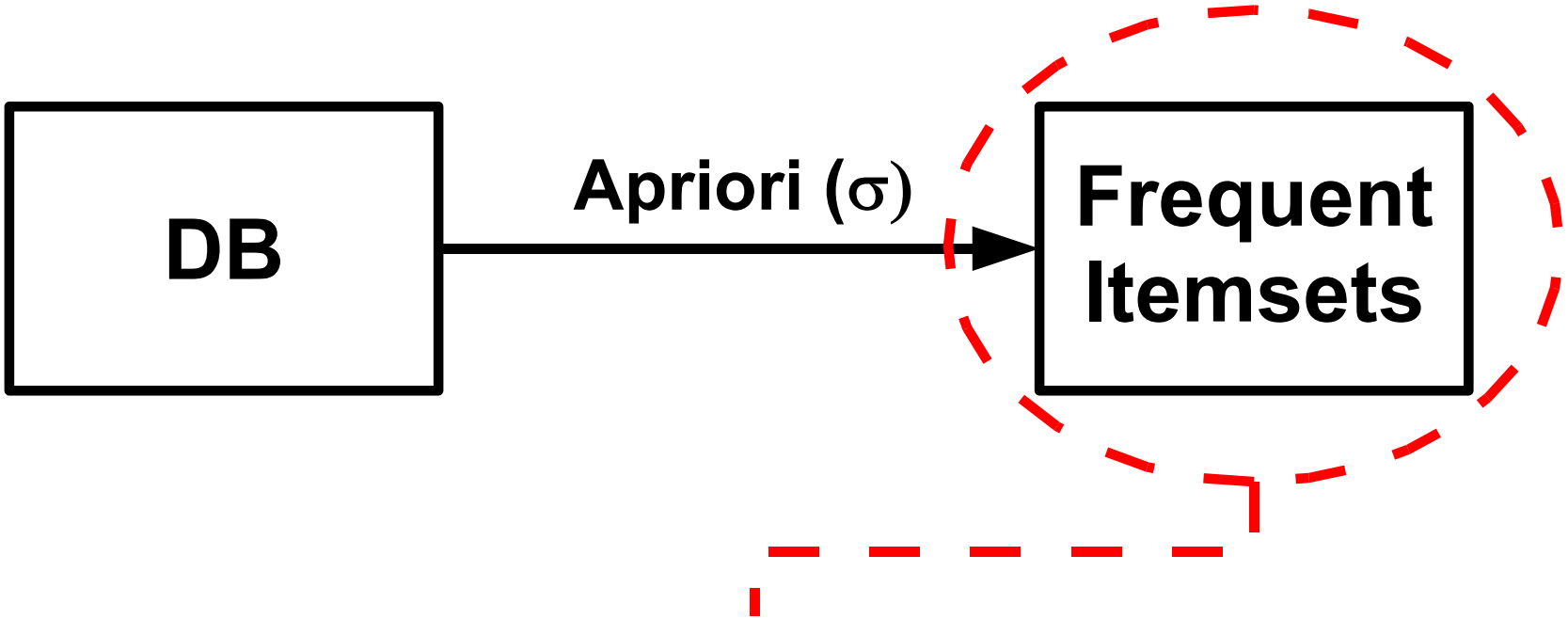
## Question

Can Dr. House publish his discoveries to third persons without offending the privacy of his patients?

# A Motivating Example in the Medical Domain



# A Motivating Example in the Medical Domain




**Does this set of itemsets violate the anonymity of individuals in DB?**

# An Association Rule Can Be Used to Break Anonymity

## Example

$$a_1 \wedge a_2 \wedge a_3 \Rightarrow a_4 \quad [sup = 80,\ conf = 98.7\%]$$

# An Association Rule Can Be Used to Break Anonymity

## Example

$$a_1 \wedge a_2 \wedge a_3 \Rightarrow a_4 \quad [sup = 80,\ conf = 98.7\%]$$

$$sup(\{a_1, a_2, a_3\}) = \frac{sup(\{a_1, a_2, a_3, a_4\})}{conf} \approx \frac{80}{0.987} = 81.05$$

# An Association Rule Can Be Used to Break Anonymity

## Example

$$a_1 \wedge a_2 \wedge a_3 \Rightarrow a_4 \quad [sup = 80,\ conf = 98.7\%]$$

$$sup(\{a_1, a_2, a_3\}) = \frac{sup(\{a_1, a_2, a_3, a_4\})}{conf} \approx \frac{80}{0.987} = 81.05$$

In other words, we know that there is just one individual for which the pattern $a_1 \wedge a_2 \wedge a_3 \wedge \neg a_4$ holds.

# Now we know that...

## Fact

- *Even if we mine with a high support value, we can infer patterns holding in the original database which are not intentionally released*

# Now we know that...

## Fact

- *Even if we mine with a high support value, we can infer patterns holding in the original database which are not intentionally released*
- *They can regards very few individuals*

# Now we know that...

## Fact

- *Even if we mine with a high support value, we can infer patterns holding in the original database which are not intentionally released*
- *They can regards very few individuals*
- *The support value of such patterns can be inferred without accessing the database*

# What do you mean with "*k*-Anonymous Pattern"?

## Definition (Anonymous Pattern)

Given a database $\mathcal{D}$ and an anonymity threshold $k$, a pattern $p$ is said to be *k-anonymous* if $sup_{\mathcal{D}}(p) \geq k$ or $sup_{\mathcal{D}}(p) = 0$.

## Definition (Inference Channel)

An Inference Channel is any set of itemsets from which it is possible to infer that a pattern $p$ is not $k$-anonymous.

We are interested in inference channels that are made of frequent itemsets.

# Example

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| T1 | a | b | c | d | e | f | g | h |
| T2 | a | b | c | d | e | | g | |
| T3 | a | b | c | d | e | | | |
| T4 | a | b | c | d | e | f | g | |
| T5 | a | b | c | d | e | | | |
| T6 | a | b | c | d | e | | | |
| T7 | a | b | | d | e | | | |
| T8 | a | | | | e | f | g | |
| T9 | | | c | d | e | f | g | |
| T10 | | | c | d | e | | | |
| T11 | | | c | d | e | f | g | h |
| T12 | a | b | | | | f | g | |

$$p = a \wedge b \wedge \neg c \wedge \neg d \wedge \neg e$$

$I = ab$
$J = abcde$

# Example

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| T1 | a | b | c | d | e | f | g | h |
| T2 | a | b | c | d | e | | g | |
| T3 | a | b | c | d | e | | | |
| T4 | a | b | c | d | e | f | g | |
| T5 | a | b | c | d | e | | | |
| T6 | a | b | c | d | e | | | |
| T7 | a | b | | d | e | | | |
| T8 | a | | | | e | f | g | |
| T9 | | | c | d | e | f | g | |
| T10 | | | c | d | e | | | |
| T11 | | | c | d | e | f | g | h |
| T12 | a | b | | | | f | g | |

$$p = a \wedge b \wedge \neg c \wedge \neg d \wedge \neg e$$

$I = ab$
$J = abcde$

# Reducing the Number of Patterns to Check

**Theorem**

$\forall p \in \mathcal{P}at(\mathcal{I}) : 0 < sup_{\mathcal{D}}(p) < k \;.\; \exists I \subseteq J \in 2^{\mathcal{I}} : \mathcal{C}_I^J.$

- Translation: we can prune the search space by looking for Inference Channels regarding only conjunctive patterns.
- This property makes possible to have a (Naïve) Inference Channel Detector Algorithm

# What do you mean with "Condensed Representation"?

## Definition (Partial Order on Inference Channels)

$\mathcal{C}_I^J \preceq \mathcal{C}_H^L$ when $I \subseteq H$ and $(J \setminus I) \subseteq (L \setminus H)$

# What do you mean with "Condensed Representation"?

## Definition (Partial Order on Inference Channels)

$\mathcal{C}_I^J \preceq \mathcal{C}_H^L$ when $I \subseteq H$ and $(J \setminus I) \subseteq (L \setminus H)$

## Example

$$\mathcal{C}_a^{ac} \preceq \mathcal{C}_{ab}^{abcd}$$

Intuitively, $a \wedge \neg c$ is less specific than $a \wedge b \wedge \neg c \wedge \neg d$, since the transactions s.t. $a \wedge \neg c$ are a superset of the transactions s.t. $a \wedge b \wedge \neg c \wedge \neg d$

# What do you mean with "Condensed Representation"?

## Definition (Partial Order on Inference Channels)

$\mathcal{C}_I^J \preceq \mathcal{C}_H^L$ when $I \subseteq H$ and $(J \setminus I) \subseteq (L \setminus H)$

## Example

$$\mathcal{C}_a^{ac} \preceq \mathcal{C}_{ab}^{abcd}$$

Intuitively, $a \wedge \neg c$ is less specific than $a \wedge b \wedge \neg c \wedge \neg d$, since the transactions s.t. $a \wedge \neg c$ are a superset of the transactions s.t. $a \wedge b \wedge \neg c \wedge \neg d$

## Definition (Maximal Inference Channel)

$\mathcal{C}_I^J$ is maximal w.r.t. $\mathcal{D}$ and $\sigma$, if $\forall \mathcal{C}_H^L \succeq \mathcal{C}_I^J$ then $sup(\mathcal{C}_H^L) = f_H^L = 0$

# Theoretical Results on Condensed Representation

## Theorem (Form of Maximal Inference Channel)

*An Inference Channel $\mathcal{C}_I^J$ is maximal iff*

1. *$I$ is closed and*
2. *$J$ is maximal*

## Theorem (Lossless Representation)

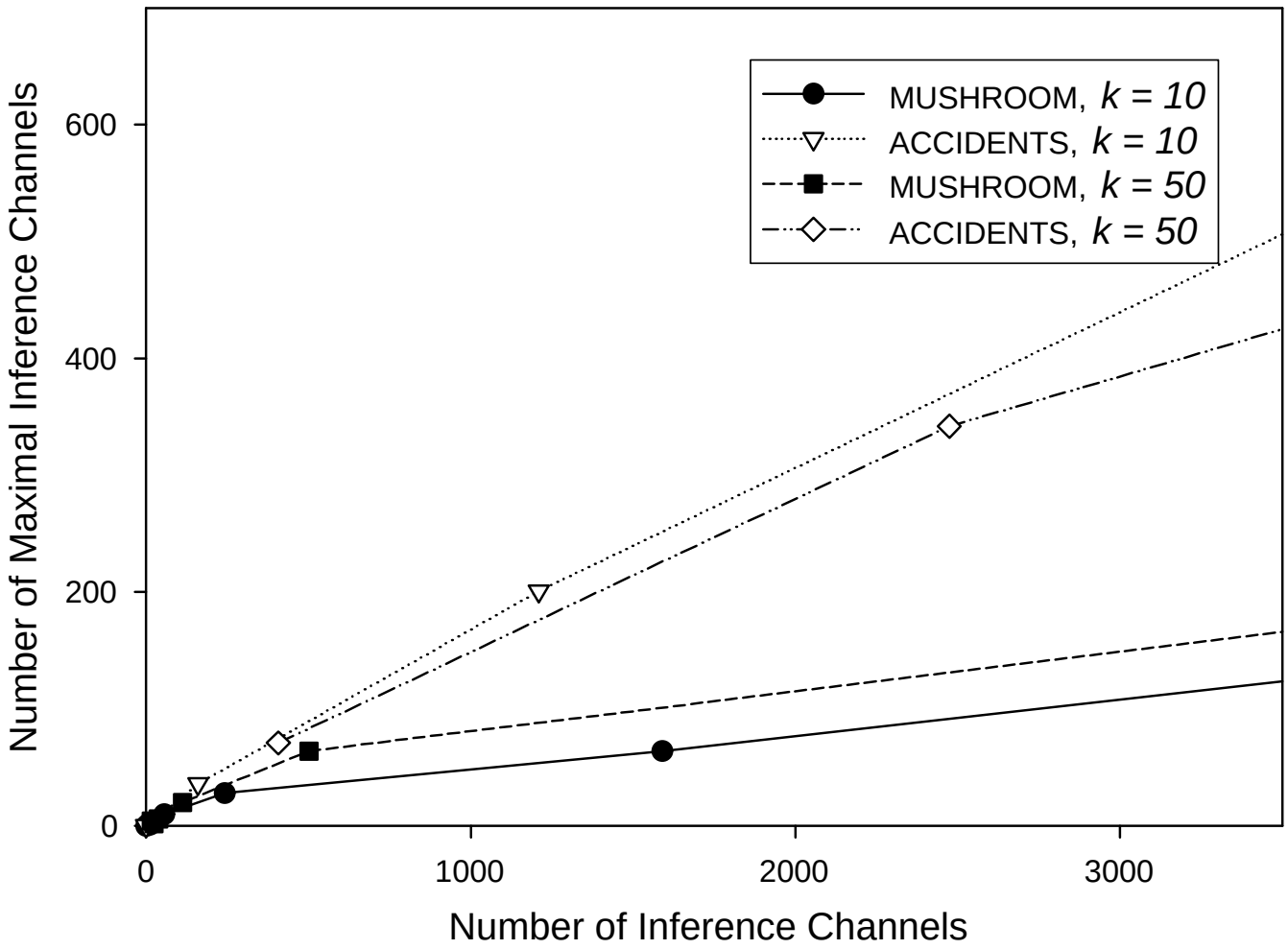
*Every Inference Channel can be computed from the set of Maximal Inference Channels*

# Condensed Representation of Inference Channels

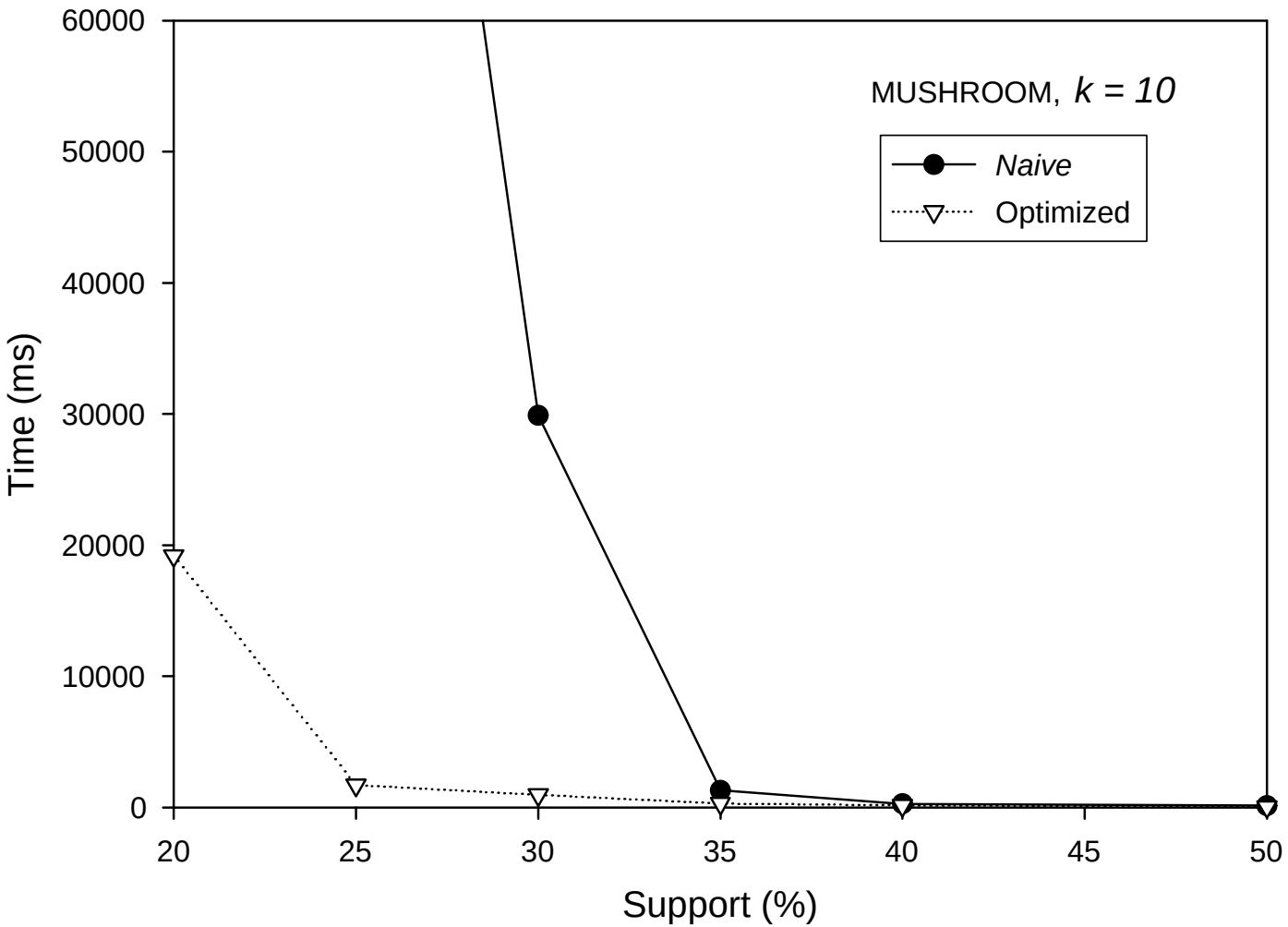
Smaller search space

1. Memory saving (the number of non-maximal channels can be huge)
2. Faster running times
3. Less distortion if we try to sanitize the set of frequent itemset (but this point is not discussed in the paper)

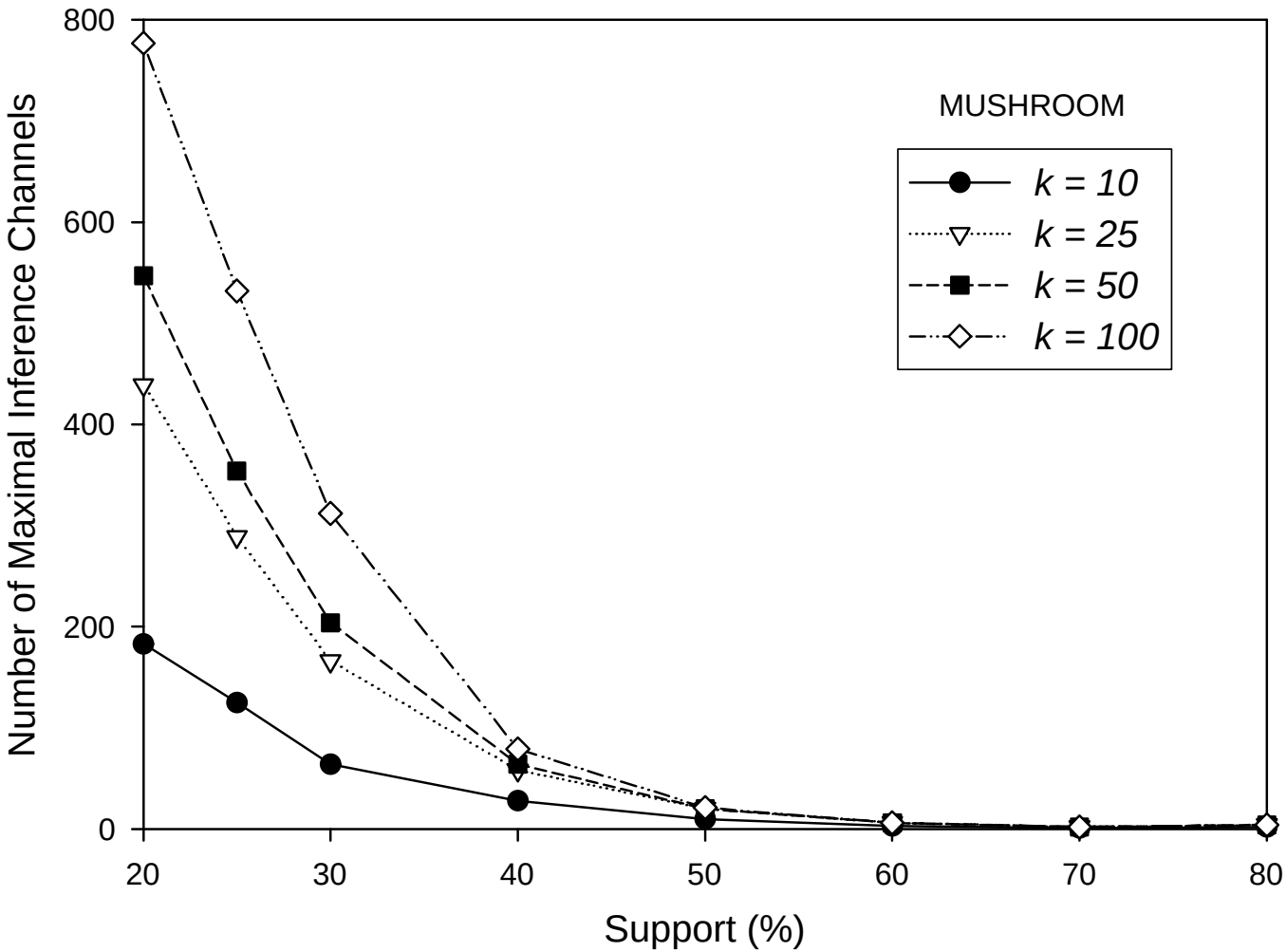
# Condensed Representation of Inference Channels



# Improvements Over the Time Performance



# Number of Inference Channels



# Summary

- We defined $k$-anonymous patterns and provide a general characterization of inference channels holding among patterns that may threat anonymity of source data
- We developed an effective and efficient algorithm to detect such potential threats, which yields a methodology to check whether the mining results may be disclosed without any risk of violating anonymity

# For Further Reading


M. Atzori, F. Bonchi, F. Giannotti, D. Pedreschi.
Blocking Anonymity Threats Raised by Frequent Itemset Mining.
*Fifth IEEE International Conference on Data Mining (ICDM)*, Houston, Texas, USA, 2005.

L. Sweeney.
k-Anonymity: A Model for Protecting Privacy.
*International Journal on Uncertainty Fuzziness and Knowledge-based Systems*, 10(5), 2002.