

November 17, 2023



Inter-University Research Institute Corporation /
Research Organization of Information and Systems

National Institute of Informatics

Supervised Pattern Mining

Data Mining 04 (データマイニング)

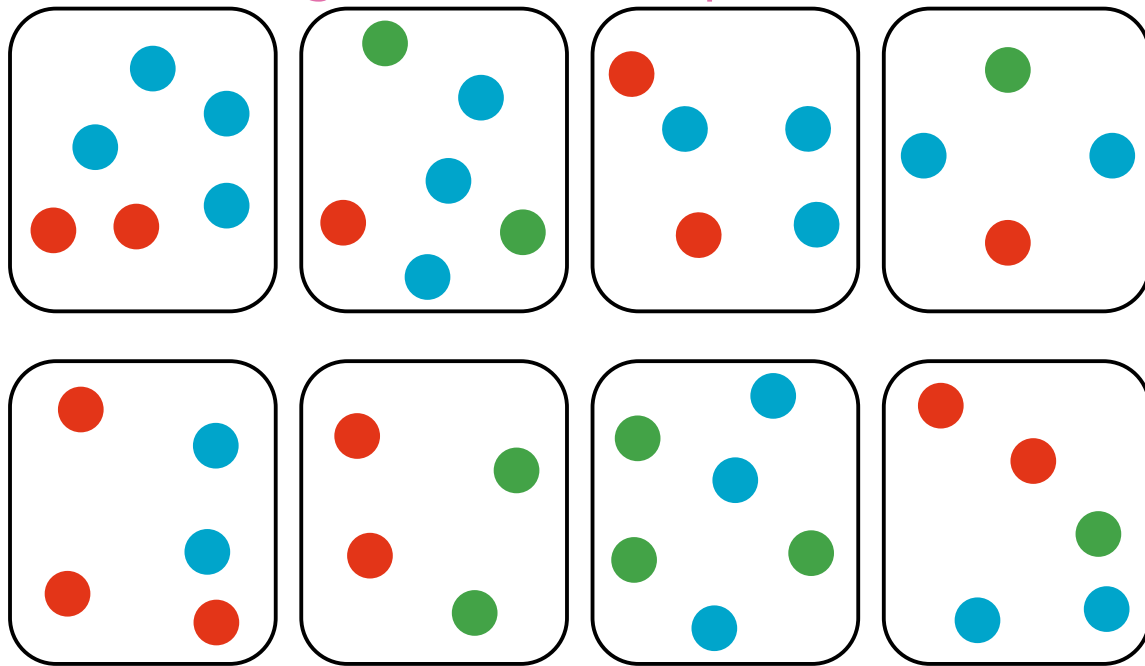
Mahito Sugiyama (杉山磨人)

Today's Outline

- Pattern mining with class labels (supervision)
 - Various measures
- Significant pattern mining
 - Statistical tests
 - Testable patterns
 - Controlling the FWER (Family-Wise Error Rate) by Tarone's testability trick

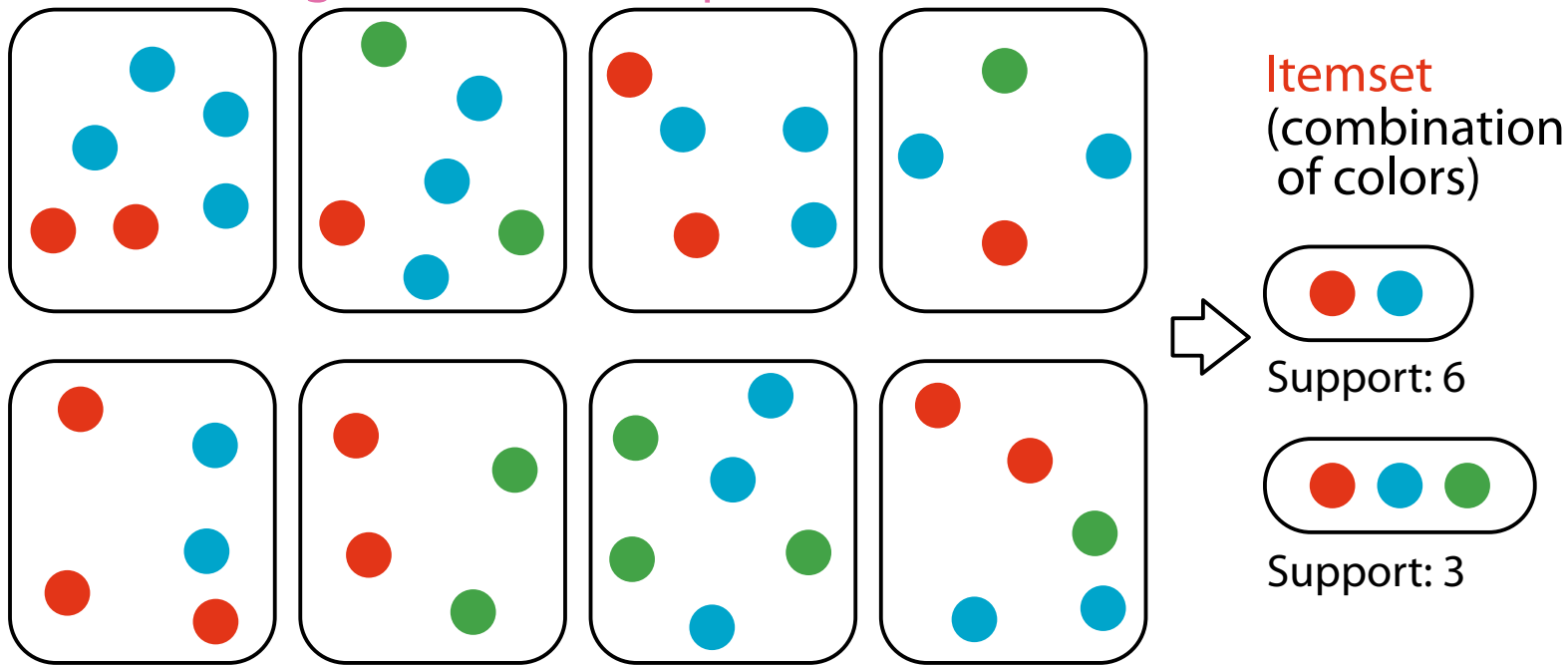
Itemset Mining

- Find **interesting combinatorial patterns** from massive data






Itemset Mining

- Find interesting combinatorial patterns from massive data

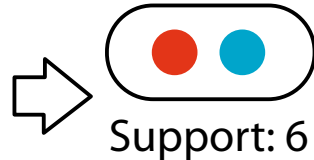


Itemset Mining (Binary Representation)

- Find interesting combinatorial patterns from massive data




			
ID1	1	1	0
ID2	1	1	1
ID3	1	1	0
ID4	1	1	1
ID5	1	1	0
ID6	0	1	1
ID7	1	0	1
ID8	1	1	1

Itemset
(combination
of colors)

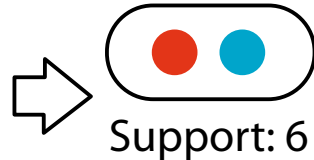


Itemset Mining (Binary Representation)

- Find interesting combinatorial patterns from massive data

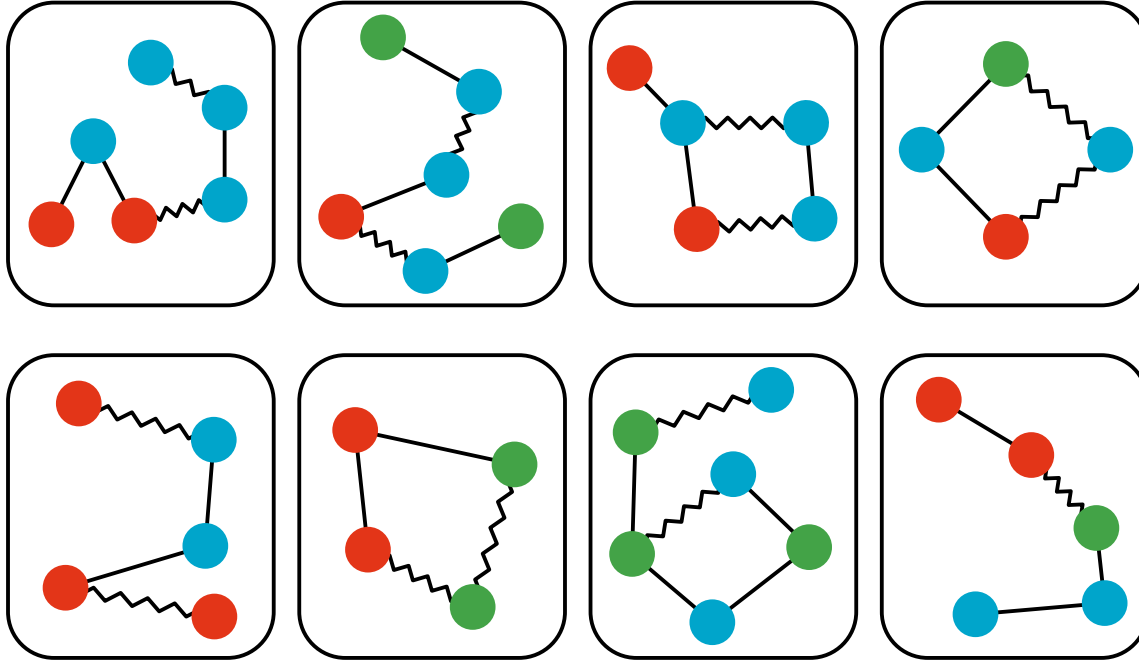
			
ID1	1	1	0
ID2	1	1	1
ID3	1	1	0
ID4	1	1	1
ID5	1	1	0
ID6	0	1	1
ID7	1	0	1
ID8	1	1	1

Itemset
(combination
of colors)



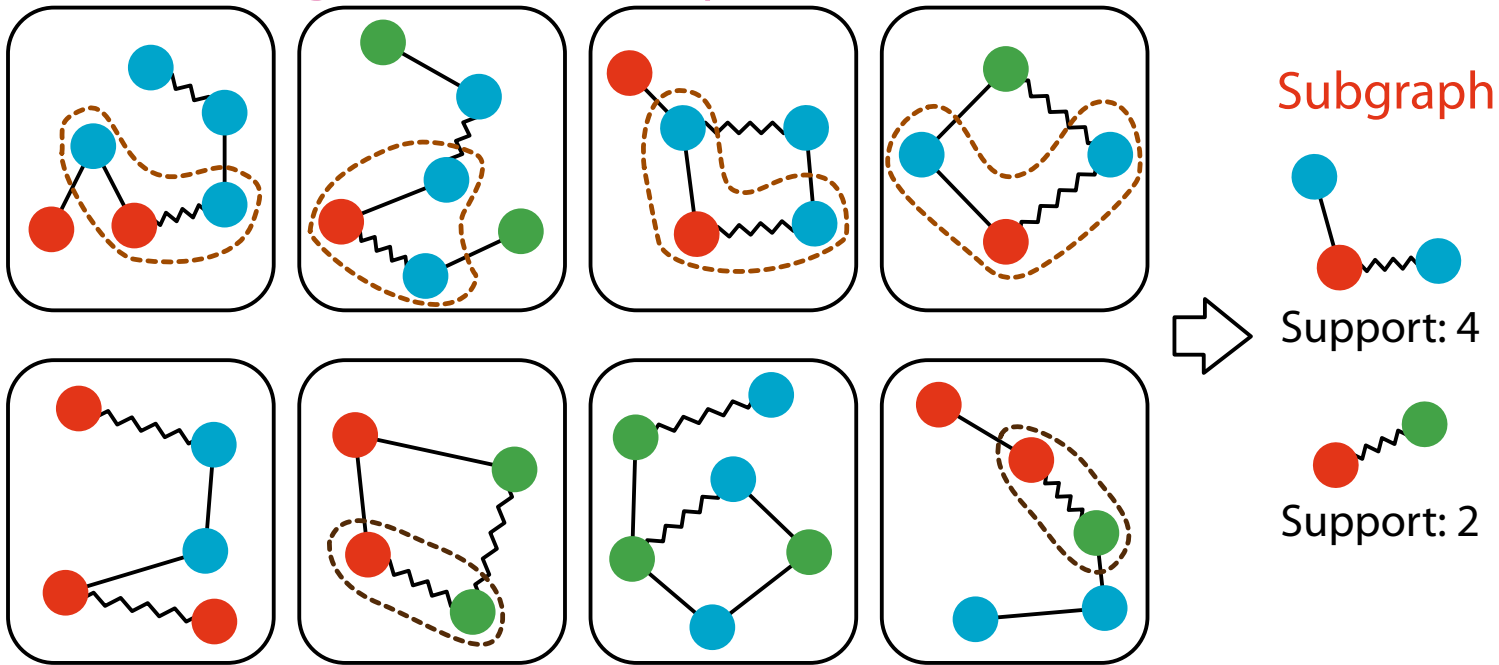
Subgraph Mining

- Find interesting combinatorial patterns from massive data



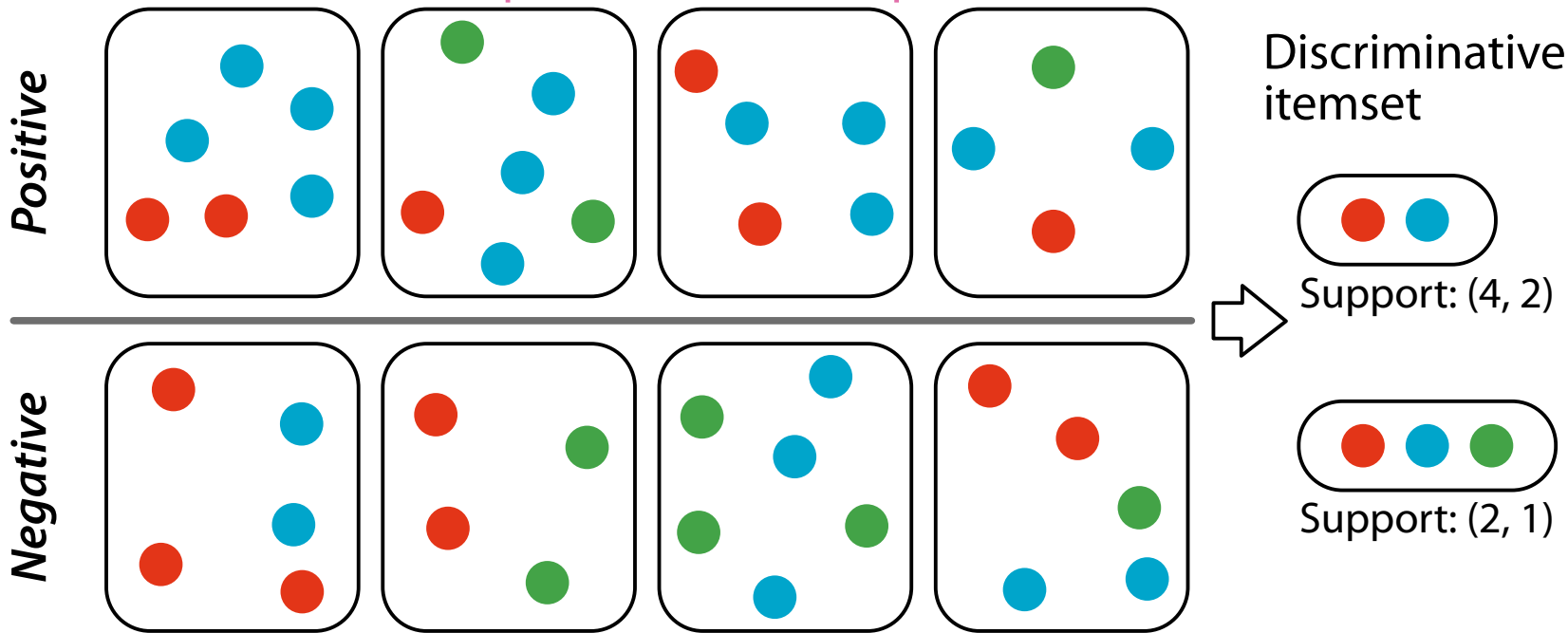
Subgraph Mining

- Find **interesting combinatorial patterns** from massive data



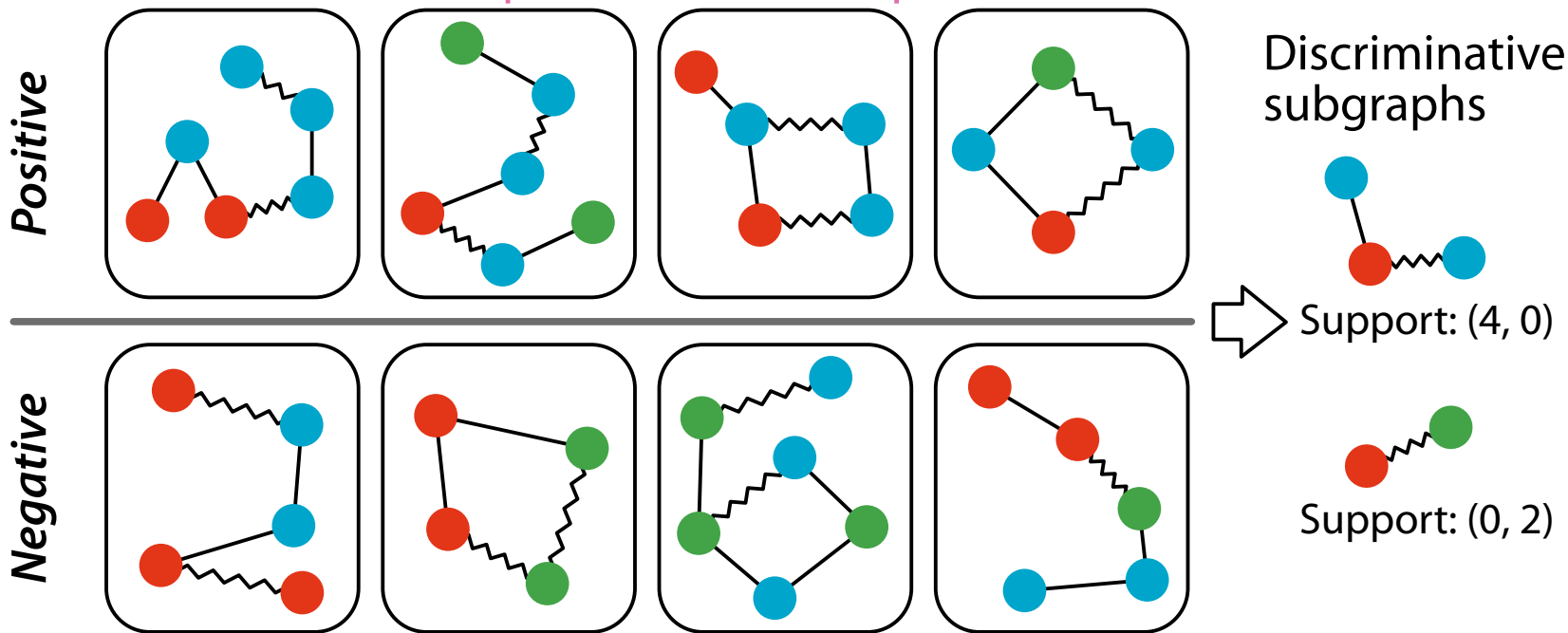
Supervised Itemset Mining

- Find **discriminative patterns** from **supervised data**



Supervised Subgraph Mining

- Find **discriminative patterns** from **supervised data**



Contingency Table

	Occurrence	Non-occurrence	Total
Positive	$\text{supp}_C(x)$	$ C - \text{supp}_C(x)$	$ C $
Negative	$\text{supp}_{\bar{C}}(x)$	$ \bar{C} - \text{supp}_{\bar{C}}(x)$	$ \bar{C} $
Total	$\text{supp}(x)$ $= \text{supp}_C(x) + \text{supp}_{\bar{C}}(x)$	$ D - \text{supp}(x)$	$ D $

Contingency Table

	Occurrence	Non-occurrence	Total
Positive	n_{11}	n_{12}	c_1
Negative	n_{21}	n_{22}	c_2
Total	s	s'	d

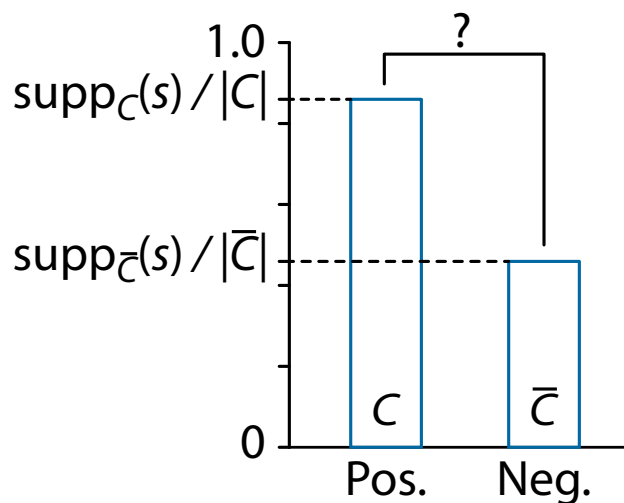
Various Measures

- Confidence: n_{11}/d
- Growth rate (relative risk): n_{11}/n_{21}
- Support difference (risk difference): $n_{11} - n_{21}$
- Mutual information:
$$\frac{n_{11}}{d} \log \frac{n_{11}/d}{c_1 s/d^2} + \frac{n_{12}}{d} \log \frac{n_{12}/d}{c_1 s'/d^2} + \frac{n_{21}}{d} \log \frac{n_{21}/d}{c_2 s/d^2} + \frac{n_{22}}{d} \log \frac{n_{22}/d}{c_2 s'/d^2}$$
- Subgroup discovery measure (weighted relative accuracy):
$$(c_1/d)((n_{11}/c_1) - (c_1/d))$$

Computing p -value of Pattern

- Given positive and negative sample sets C, \bar{C} such that $D = C \cup \bar{C}$
- The p -value of each pattern s is assessed by the Fisher's exact test

	Occ.	Non-occ.	Total
C (Pos.)	$\text{supp}_C(s)$	$ C - \text{supp}_C(s)$	$ C $
\bar{C} (Neg.)	$\text{supp}_{\bar{C}}(s)$	$ \bar{C} - \text{supp}_{\bar{C}}(s)$	$ \bar{C} $
D (Total)	$\text{supp}(s)$	$ D - \text{supp}(s)$	$ D $

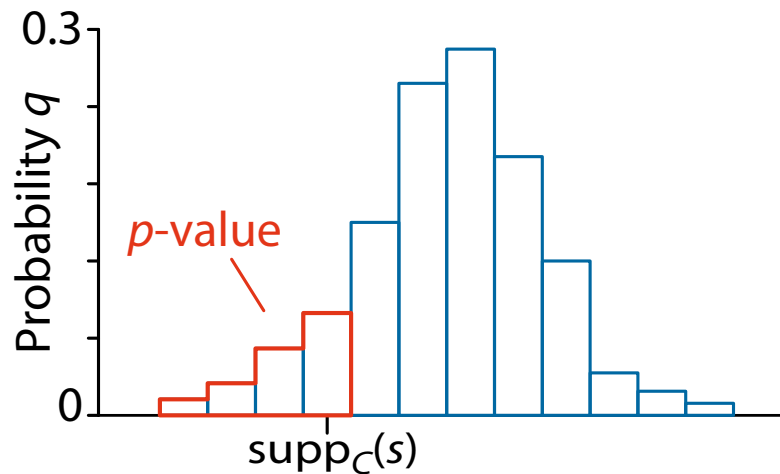


Fisher's Exact Test

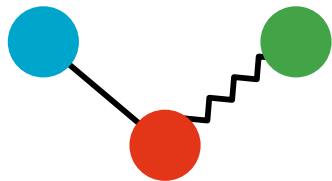
- Probability $q(\text{supp}_C(s))$ is given by hypergeometric distribution:

$$q(\text{supp}_C(s)) = \frac{\binom{|C|}{\text{supp}_C(s)} \binom{|\bar{C}|}{\text{supp}_{\bar{C}}(s)}}{\binom{|D|}{\text{supp}(s)}}$$

	Occ.	Non-occ.	Total
C (Pos.)	$\text{supp}_C(s)$	$ C - \text{supp}_C(s)$	$ C $
\bar{C} (Neg.)	$\text{supp}_{\bar{C}}(s)$	$ \bar{C} - \text{supp}_{\bar{C}}(s)$	$ \bar{C} $
D (Total)	$\text{supp}(s)$	$ D - \text{supp}(s)$	$ D $



Hypothesis Test for Each Pattern



Alternative hypothesis
is true

Null hypothesis
is true

Declared significant
($p\text{-value} < \alpha$)

True Positive

False Positive
(Type I Error)

Declared
non-significant

False Negative
(Type II Error)

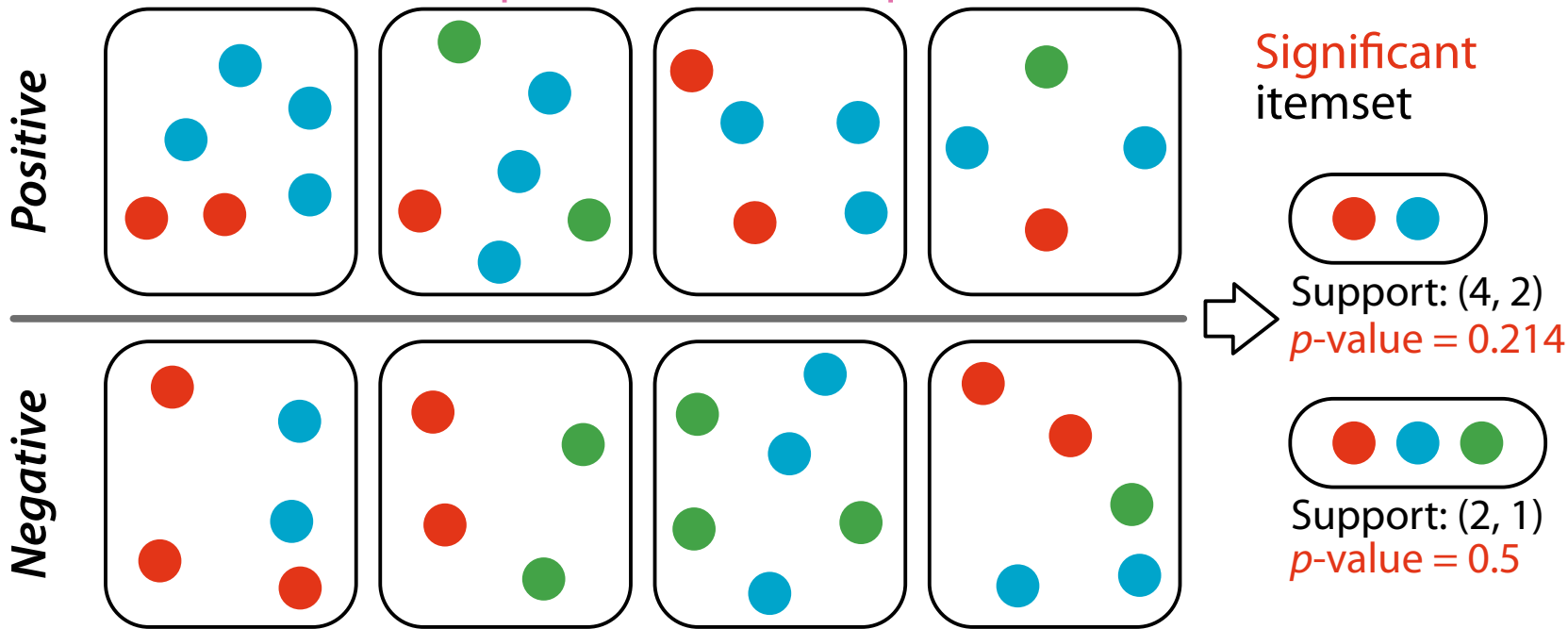
True Negative

Null: Occurrence of pattern is **independent** from classes

Alternative: Occurrence of pattern is **associated with** classes

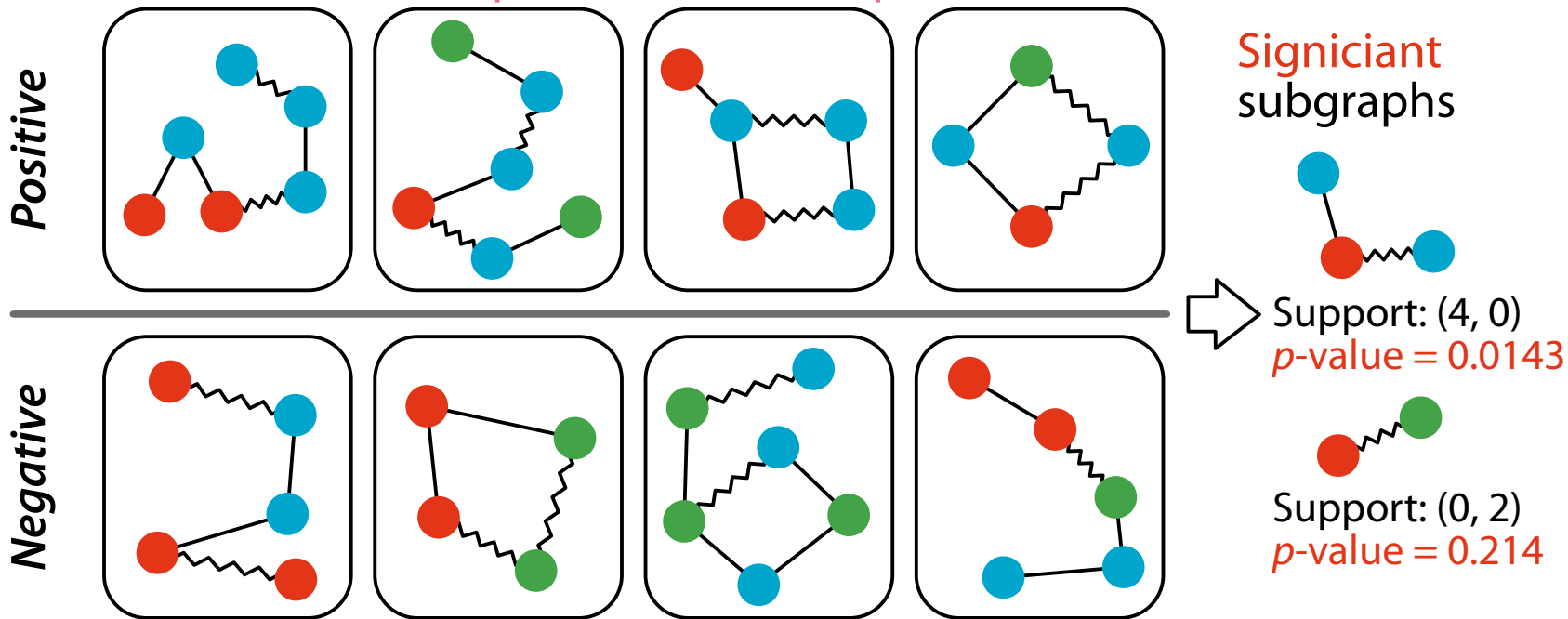
Significant Pattern (Itemset) Mining

- Find **discriminative patterns** from **supervised data**



Significant Subgraph Mining

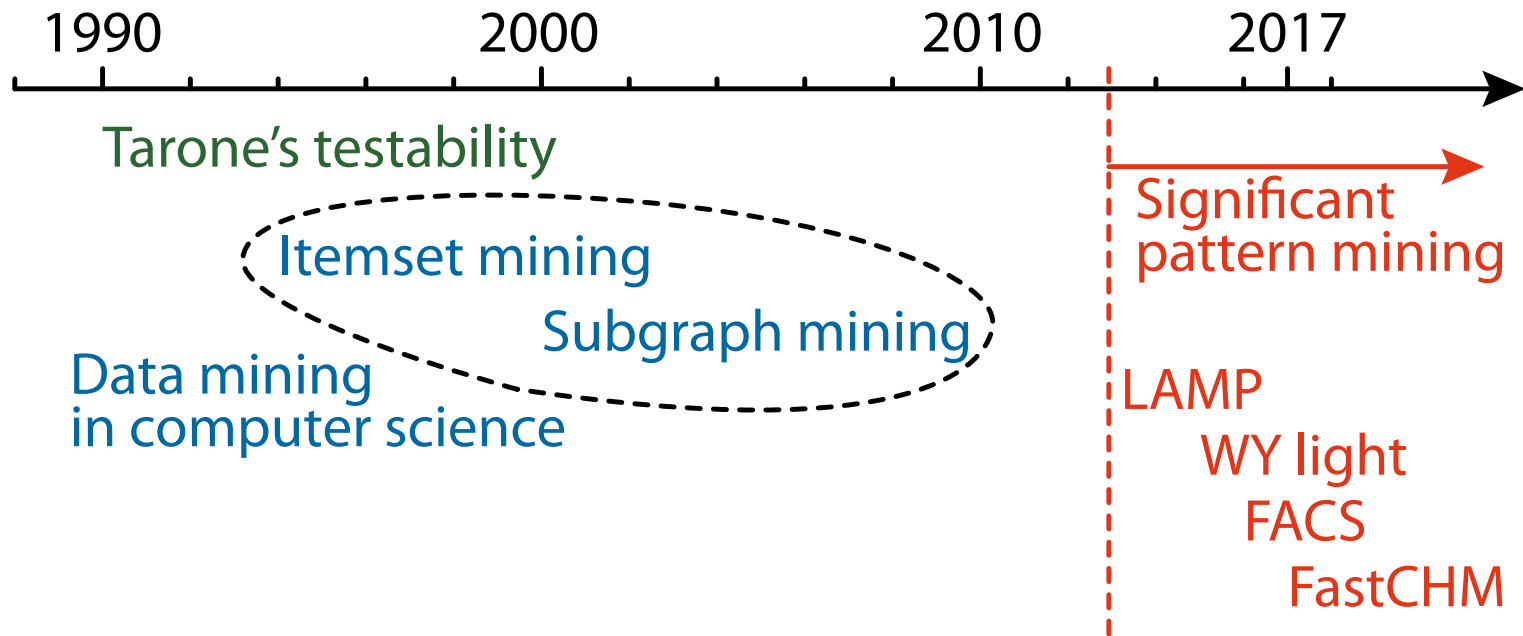
- Find **discriminative patterns** from **supervised data**



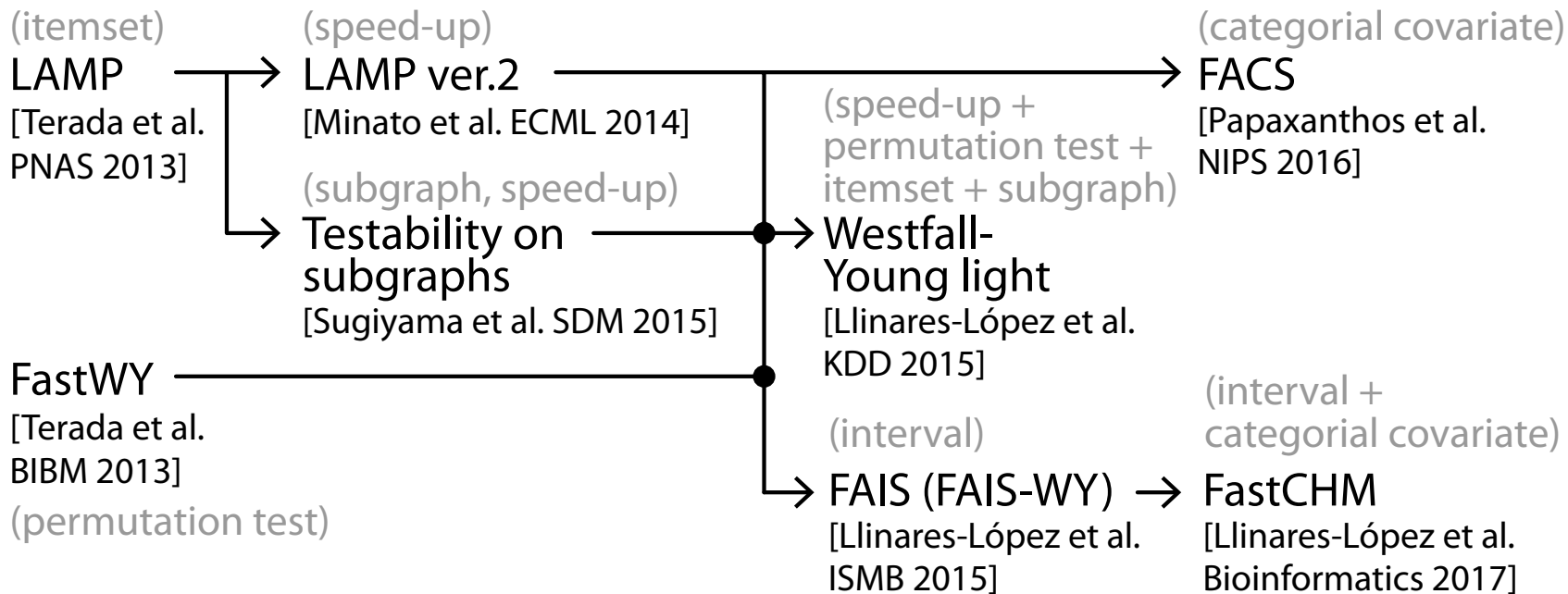
Challenges and Solutions of SPM

1. **(Computational)** How to check all patterns with avoiding **combinatorial explosion**?
2. **(Statistical)** How to measure the statistical association (i.e. p -value) with correcting for multiple testing with avoiding **combinatorial explosion**?
 - **Answer: Tarone's trick + Apriori principle**
 - The **Tarone's trick** to define patterns that are irrelevant
 - The **Apriori principle** to efficiently prune such patterns using the partial order structure of patterns

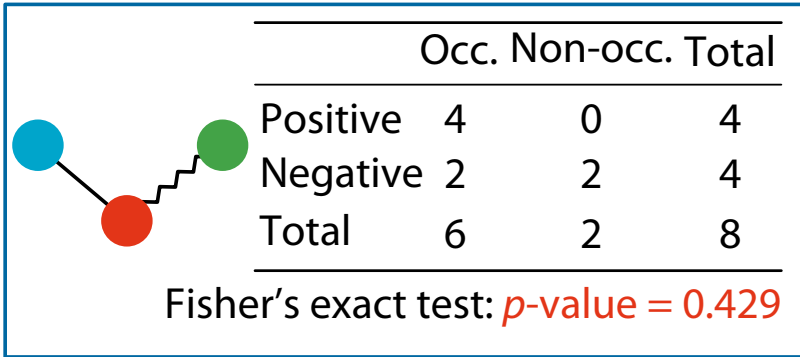
Timeline



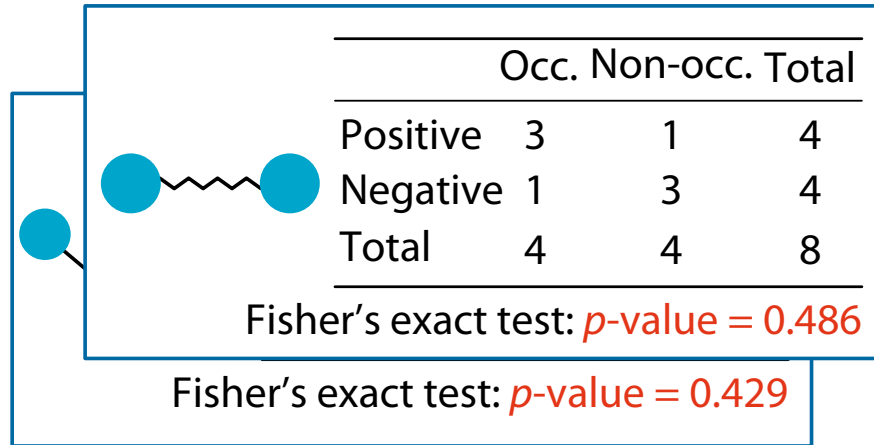
Summary of SPM Methods



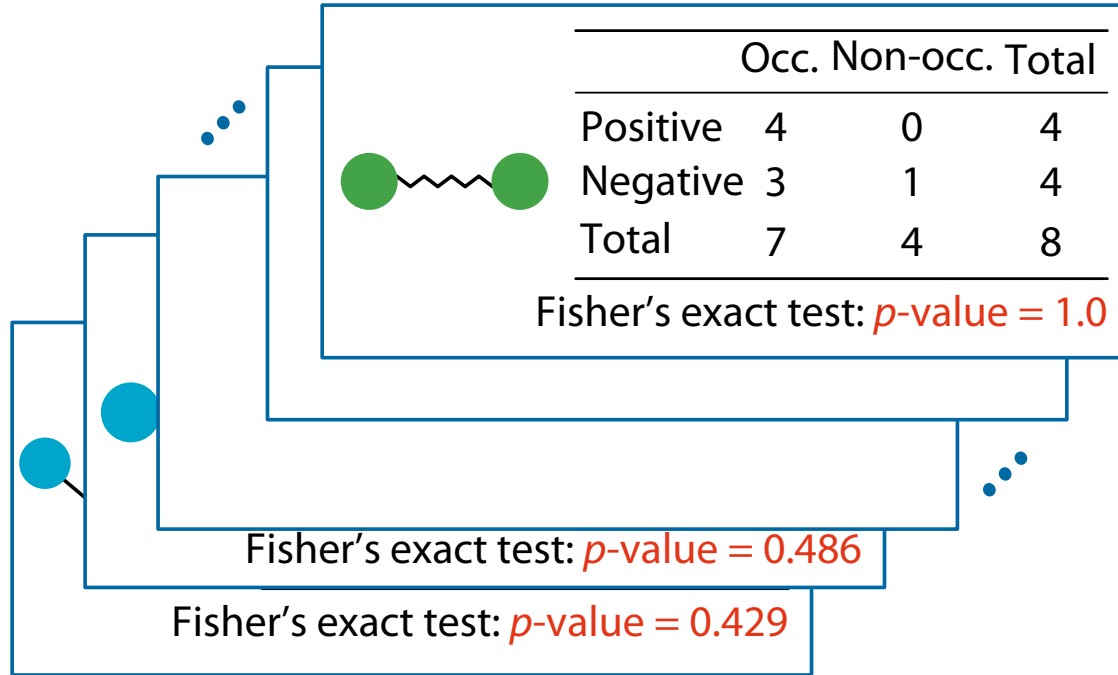
Multiple Testing



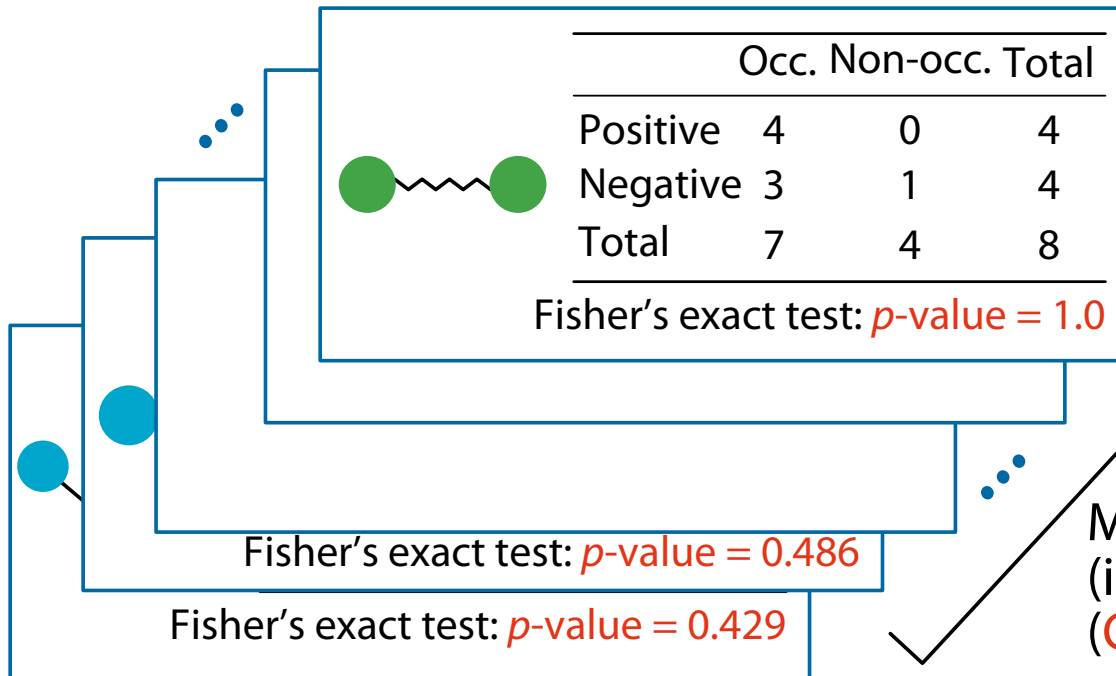
Multiple Testing



Multiple Testing



Multiple Testing



Task: Enumerate **all significant patterns** while controlling the **FWER**

Massive number of (infinitely many) patterns (**Combinatorial explosion!**)

Multiple Testing Correction

- In each test, [probability of having a false positive] $\leq \alpha$
- If we repeat m tests, αm patterns can be false positives
 - Too many if m is large! For example in itemset mining:
 - For 100000 items, #patterns = 2^{100000}
 - Set significance level $\alpha = 0.01$
 - Number of false positives: $0.01 \cdot 2^{100000} = 10^{30101}$

Multiple Testing Correction

- In each test, [probability of having a false positive] $\leq \alpha$
- If we repeat m tests, αm patterns can be false positives
 - Too many if m is large! For example in itemset mining:
 - For 100000 items, #patterns = 2^{100000}
 - Set significance level $\alpha = 0.01$
 - Number of false positives: $0.01 \cdot 2^{100000} = 10^{30101}$
- **FWER** (family-wise error rate): ***Probability of having more than one false positives among all patterns***
 - $\text{FWER} = 1 - (1 - \alpha)^m$ if patterns are independent

Controlling the FWER

- $\text{FWER} = \Pr(\text{FP} > 0)$
 - FP: Number of false positives
- To achieve $\text{FWER} = \alpha$, change the significance level for each pattern
from α to δ ($\delta \leq \alpha$), the corrected significance level

Controlling the FWER

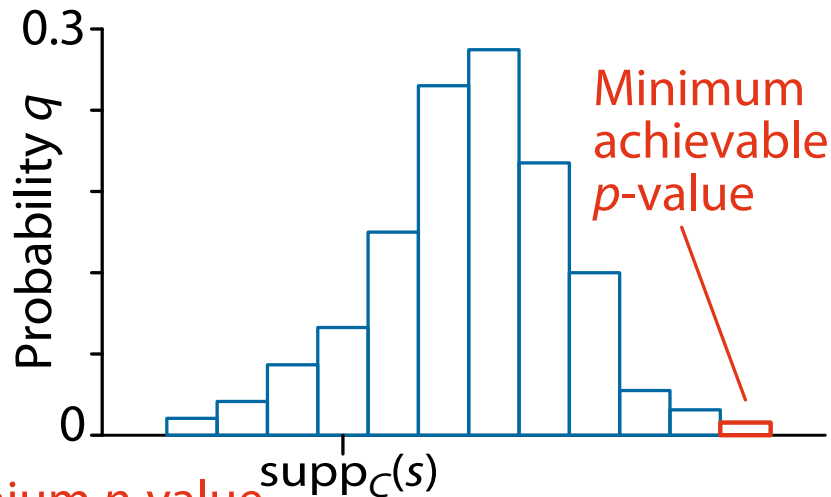
- $\text{FWER} = \Pr(\text{FP} > 0)$
 - FP: Number of false positives
- To achieve $\text{FWER} = \alpha$, change the significance level for each pattern from α to δ ($\delta \leq \alpha$), the **corrected significance level**
- **Objective:** **Maximize $\text{FWER}(\delta)$ subject to $\text{FWER}(\delta) \leq \alpha$**
 - $\text{FWER}(\delta)$: FWER at corrected significance level δ
 - Cannot be evaluated in closed form (simple but not easy!)
 - **Bonferroni correction** is popular: $\delta_{\text{Bon}}^* = \alpha/m$

Minimum Achievable p -value $\Psi(\sigma)$

- Consider the minimum achievable p -value $\Psi(s)$ of a pattern s for its support $\text{supp}(s)$

	Occ.	Non-occ.	Total
C (Pos.)	$\text{supp}(s)$	$ C - \text{supp}_C(s)$	$ C $
\bar{C} (Neg.)	0	$ \bar{C} - \text{supp}_{\bar{C}}(s)$	$ \bar{C} $
D (Total)	$\text{supp}(s)$	$ D - \text{supp}(s)$	$ D $

Most biased case that achieves the minimum p -value

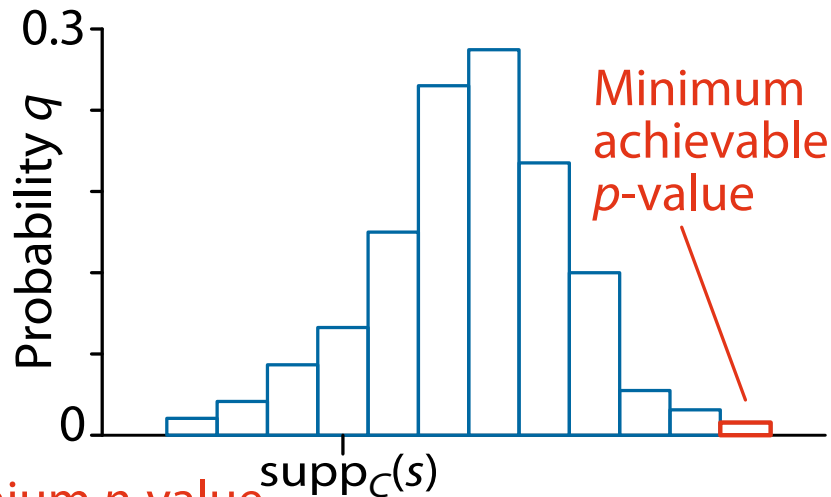


Computing $\Psi(s)$

Minimum achievable p -value $\Psi(s) = \frac{\binom{|C|}{\text{supp}(s)}}{\binom{|D|}{\text{supp}(s)}}$

	Occ.	Non-occ.	Total
C (Pos.)	$\text{supp}(s)$	$ C - \text{supp}_C(s)$	$ C $
\bar{C} (Neg.)	0	$ \bar{C} - \text{supp}_{\bar{C}}(s)$	$ \bar{C} $
D (Total)	$\text{supp}(s)$	$ D - \text{supp}(s)$	$ D $

Most biased case that achieves the minimum p -value



Tarone's Testability Trick

$$\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots, \mathcal{F}_{m-1}, \mathcal{F}_m, \mathcal{F}_{m+1}, \dots, \mathcal{F}_{2^d} \quad (\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1}))$$

Tarone's Testability Trick

$$m \psi(\mathcal{F}_m) < \alpha \quad \text{and} \quad (m+1)\psi(\mathcal{F}_{m+1}) \geq \alpha$$

$$\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots, \mathcal{F}_{m-1}, \mathcal{F}_m, \mathcal{F}_{m+1}, \dots, \mathcal{F}_{2^d} \quad (\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1}))$$

Tarone's Testability Trick

$$m \psi(\mathcal{F}_m) < \alpha \quad \text{and} \quad (m+1)\psi(\mathcal{F}_{m+1}) \geq \alpha$$

$\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots, \mathcal{F}_{m-1}, \mathcal{F}_m, \mathcal{F}_{m+1}, \dots, \mathcal{F}_{2^d}$ $(\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1}))$

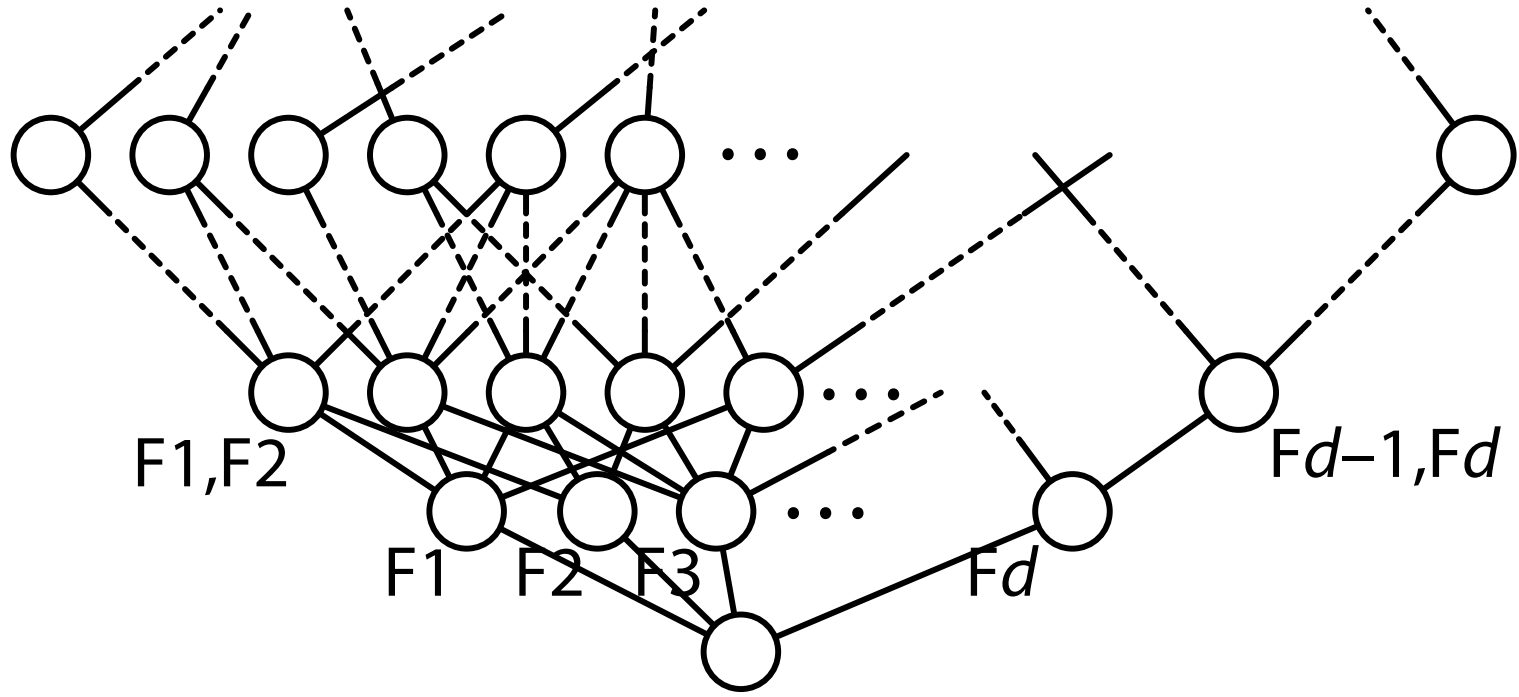
Testable
patterns

Untestable
patterns

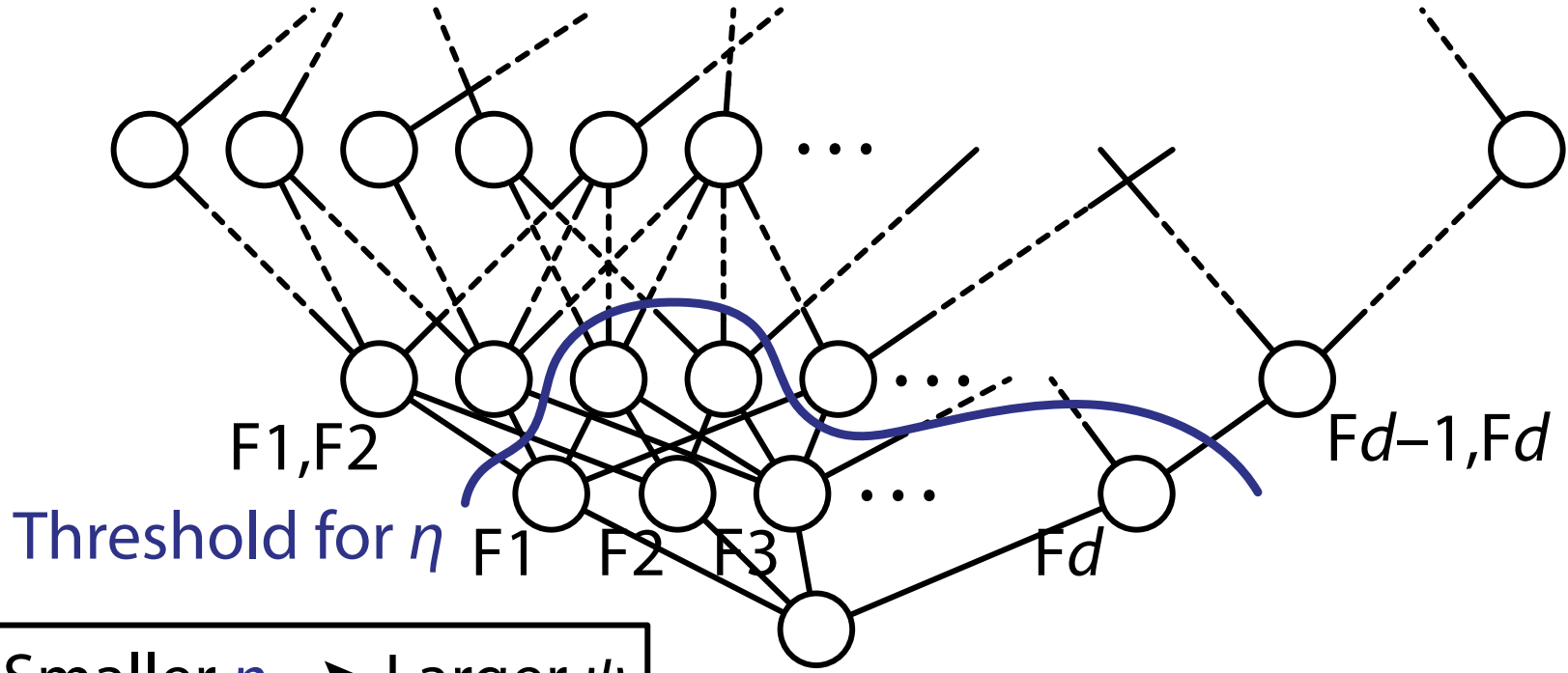
→ Prune without testing

↓
 \mathcal{F}_i is significant if: $p\text{-value}(\mathcal{F}_i) < \alpha / \textcircled{m}$ — Correction factor

Enumeration Algorithm Based on Apriori

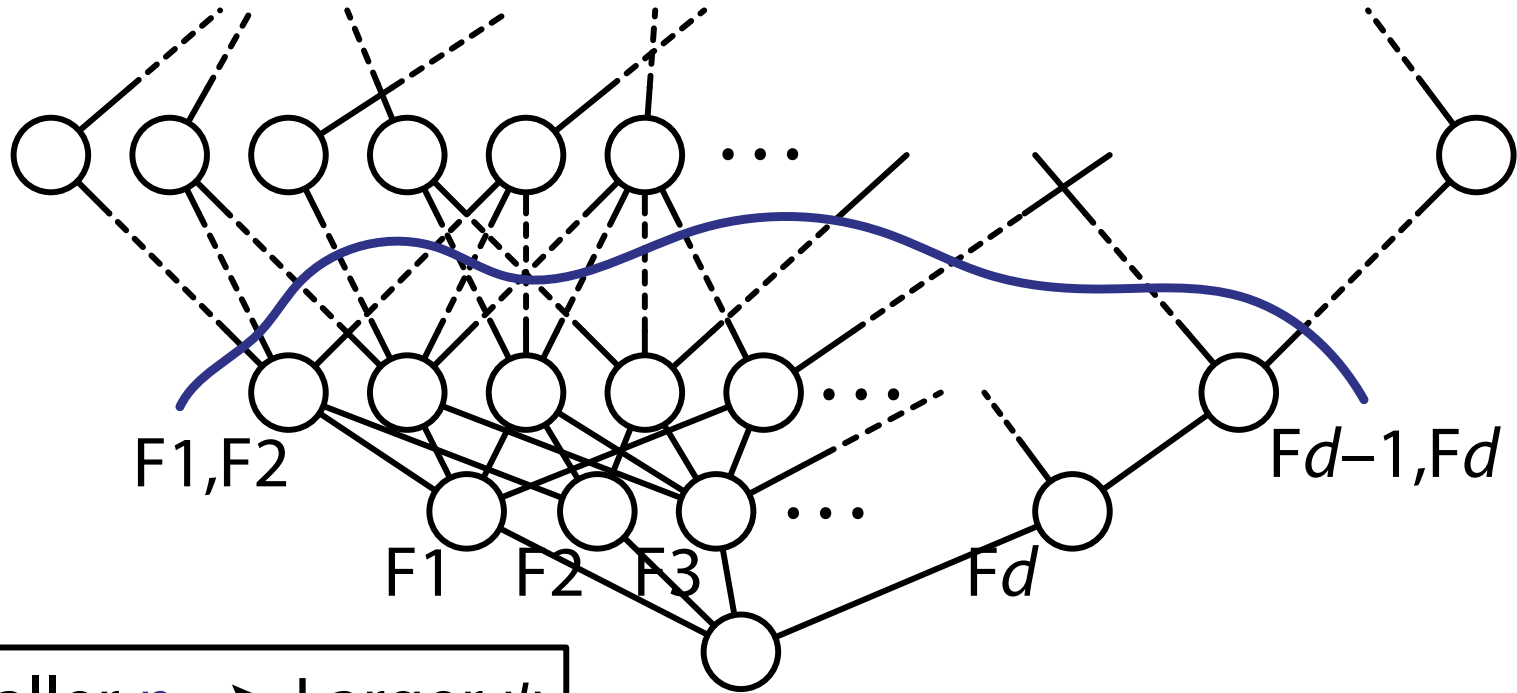


Enumeration Algorithm Based on Apriori



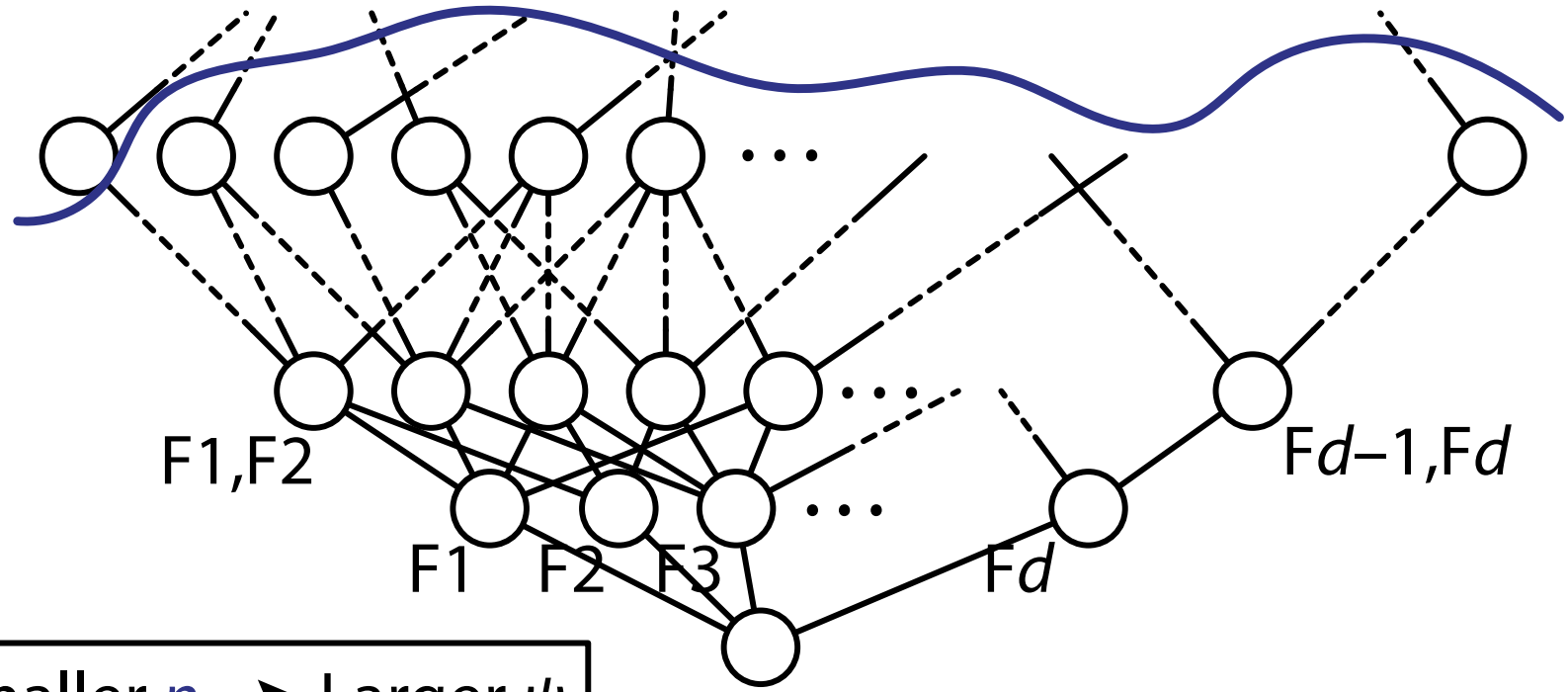
Smaller $\eta \rightarrow$ Larger ψ

Enumeration Algorithm Based on Apriori



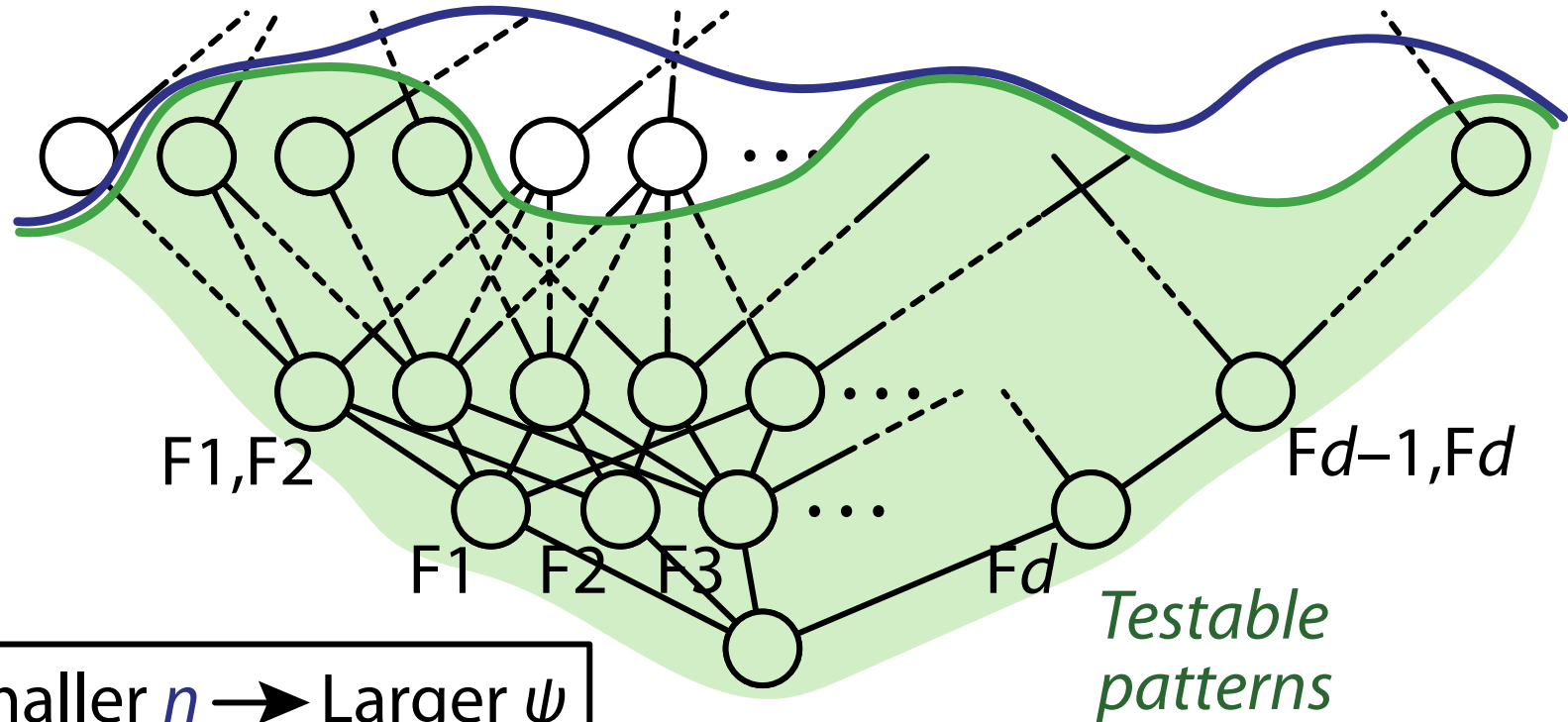
Smaller $\eta \rightarrow$ Larger ψ

Enumeration Algorithm Based on Apriori



Smaller $\eta \rightarrow$ Larger ψ

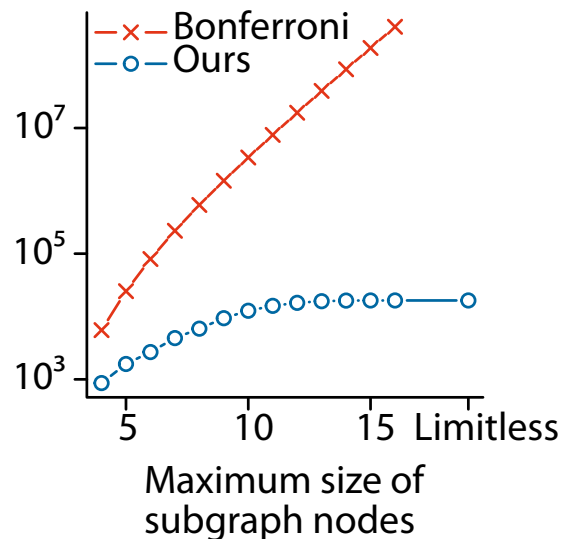
Enumeration Algorithm Based on Apriori



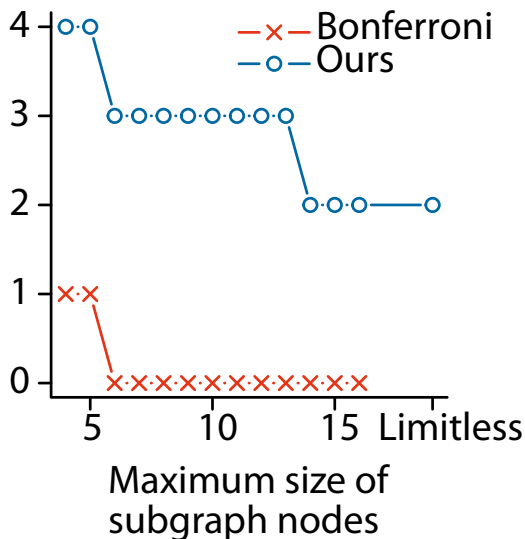
Smaller $\eta \rightarrow$ Larger ψ

Power of Testability (Subgraph Mining)

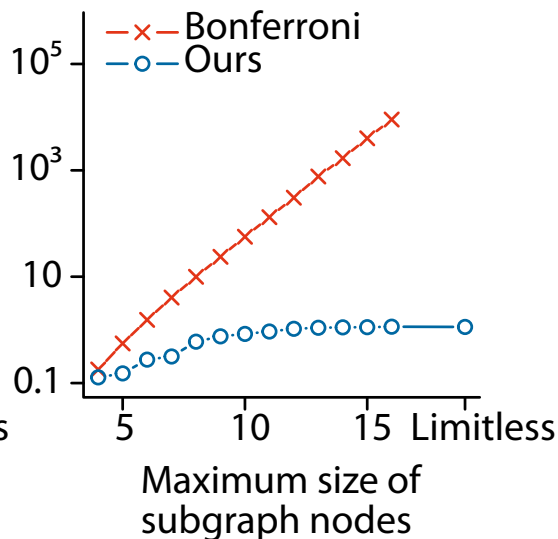
Correction factor



Number of significant subgraphs



Running time (second)



The PTC (Predictive Toxicology Challenge) dataset with 601 chemical compounds

Conclusion

- Supervised pattern mining has been considered
- Significant pattern mining is introduced
 - Find statistically significant patterns while controlling the FWER
 - pattern mining (data mining) + multiple testing correction (statistics)
 - Key technique: Tarone's trick