August 3, 2015

DOI: 10.1145/2783258.2783363

ERATO 感謝祭
Season II

Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing (KDD2015)

Felipe Llinares-López, Mahito Sugiyama,
Laetitia Papaxanthos, Karsten Borgwardt (ETH Zürich)

杉山 麿人 (大阪大学，さきがけ研究者)

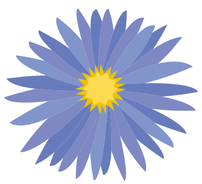
Summary

- Computing *p*-values in (supervised) pattern mining
  - Itemsets, subgraphs, ...
  - Significant pattern mining
- **Challenge:** How to correct for multiple testing?
  - Control the false positive rate of resulting patterns
  - Number of patterns are massive (more than billions!)
- We propose a new method "Westfall-Young light"
  - Empirically estimate the null distribution of pattern frequencies in each class via permutations
  - Embed "permutation + *p*-value computation" into pattern mining

Itemset Mining (GWAS)

Items (SNPs)

Case



Sample 1: 0 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 1 1 1 0

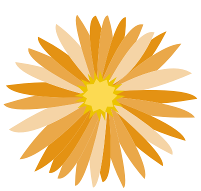
Sample 2: 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 1 0 1 0 0

Sample 3: 1 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1

Sample 4: 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1

Sample 5: 1 1 0 1 1 0 1 1 1 0 0 1 0 1 0 1 0 0 0 0

Control



Sample 6: 0 0 1 1 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0 0

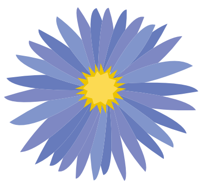
Sample 7: 0 1 0 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0

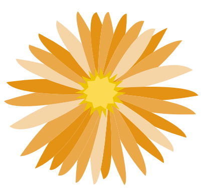
Sample 8: 1 0 1 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 0 0

Sample 9: 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1

Itemset Mining (GWAS)

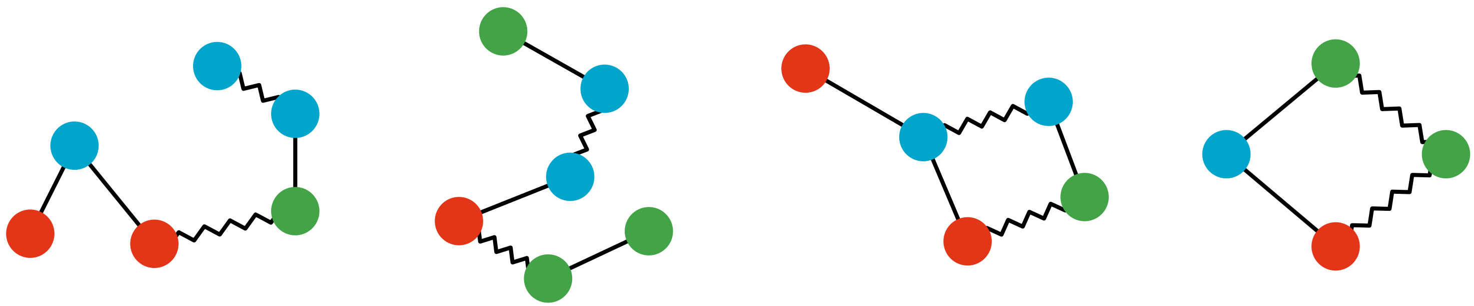
		Items (SNPs)
Case	Sample 1:	0 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 1 1 1 0
	Sample 2:	1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 1 0 1 0 0
	Sample 3:	1 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1
	Sample 4:	1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1
	Sample 5:	1 1 0 1 1 0 1 1 1 0 0 1 0 1 0 1 0 0 0 0
Control	Sample 6:	0 0 1 1 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0 0
	Sample 7:	0 1 0 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0
	Sample 8:	1 0 1 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 0 0
	Sample 9:	1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1



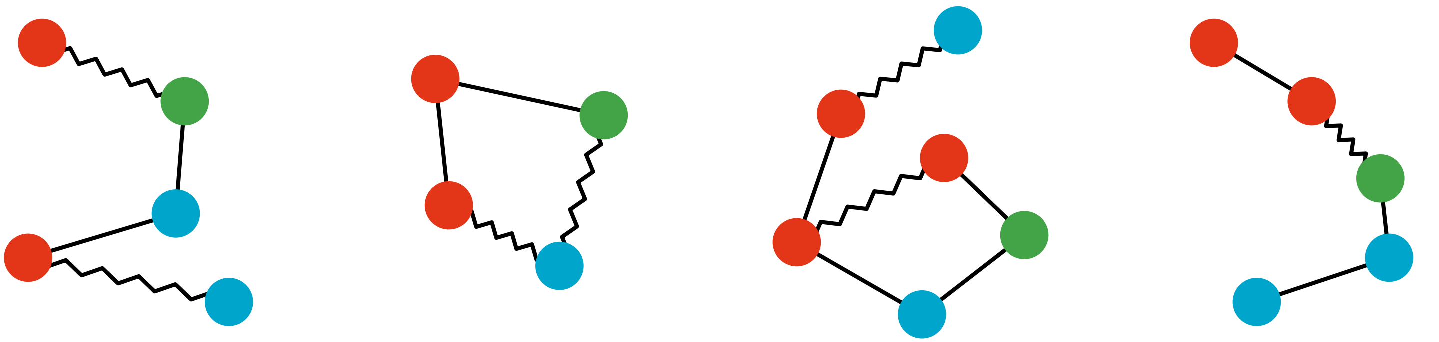


Subgraph Mining (Drag Discovery)

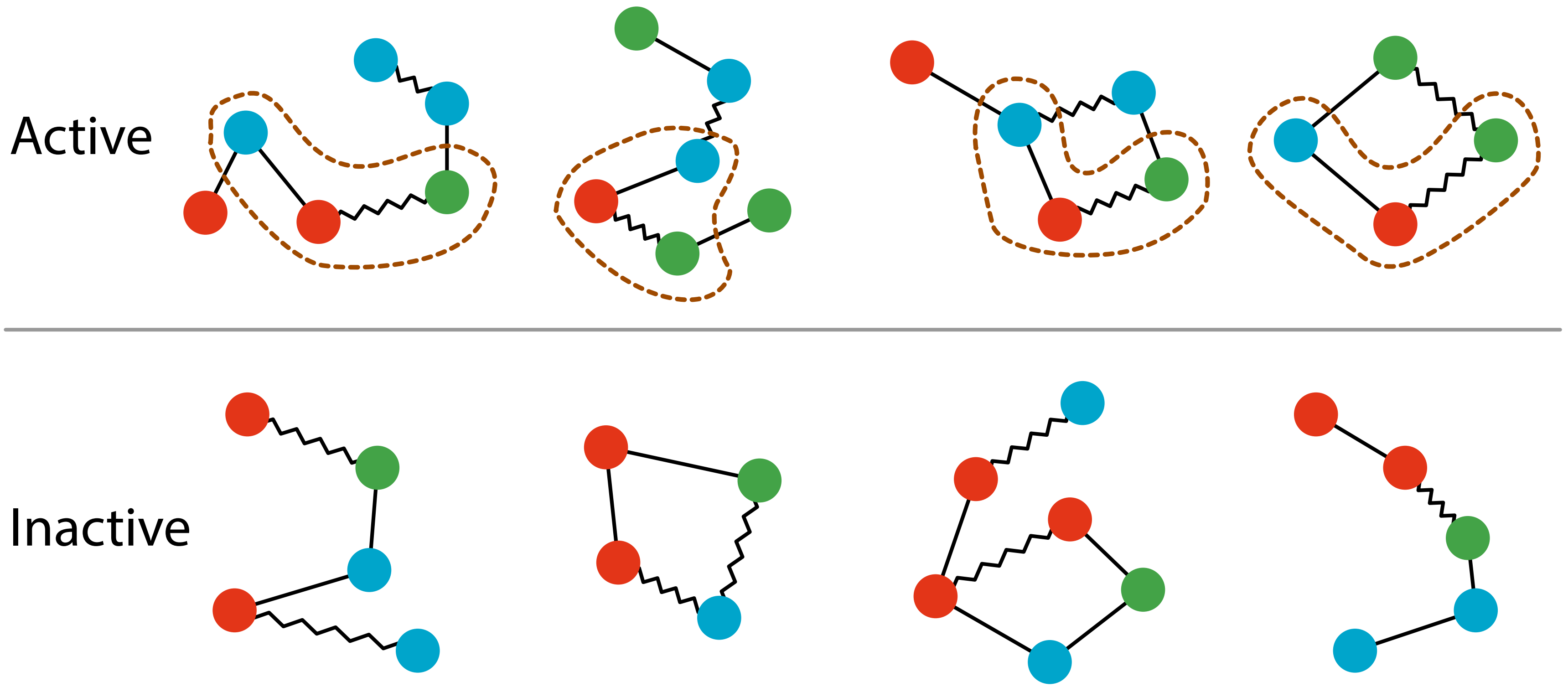
Active



Inactive



Subgraph Mining (Drag Discovery)



Timeline

(Itemset)
LAMP
[Terada et al. PNAS 2013]

(Speed-up)
LAMP ver.2
[Minato et al. ECML 2014]

(Subgraph, Speed-up)
Testability on subgraphs
[Sugiyama et al. SDM 2015]

FastWY
[Terada et al. BIBM 2013]
(Permutation test)

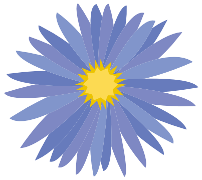
(Speed-up + Permutation test + Itemset + Subgraph)
Westfall-Young light
[Llinares-López et al. KDD 2015]

(Interval)
FAIS (FAIS-WY)
[Llinares-López et al. ISMB 2015]

Itemset Mining (GWAS)

Items (SNPs)

Case



Sample 1: 0 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 1 1 1 0

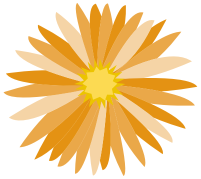
Sample 2: 1 1 0 1 1 0 1 1 1 0 0 0 0 1 0 1 0 1 0 0

Sample 3: 1 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1

Sample 4: 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1

Sample 5: 1 1 0 1 1 0 1 1 1 0 0 1 0 1 0 1 0 0 0 0

Control



Sample 6: 0 0 1 1 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0 0

Sample 7: 0 1 0 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0

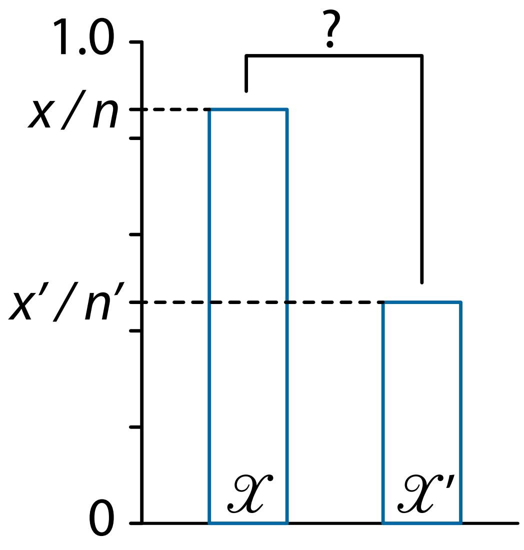
Sample 8: 1 0 1 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 0 0

Sample 9: 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1

Testing the Independence of Pattern

- Given two sets of transactions $\mathcal{X}, \mathcal{X}'$
  - $|\mathcal{X}| = n, |\mathcal{X}'| = n'$ $(n \le n')$
- The *p*-value of each pattern (itemset) H is determined by the Fisher's exact test
  - $x = |\{ X \in \mathcal{X} \mid H \subseteq X \}|$

	Occ.	Non-occ.	Total
$\mathscr{X}$	x	$n - x$	n
$\mathscr{X}'$	x'	$n' - x'$	n'
Total	$x + x'$	$(n - x) + (n' - x')$	$n + n'$

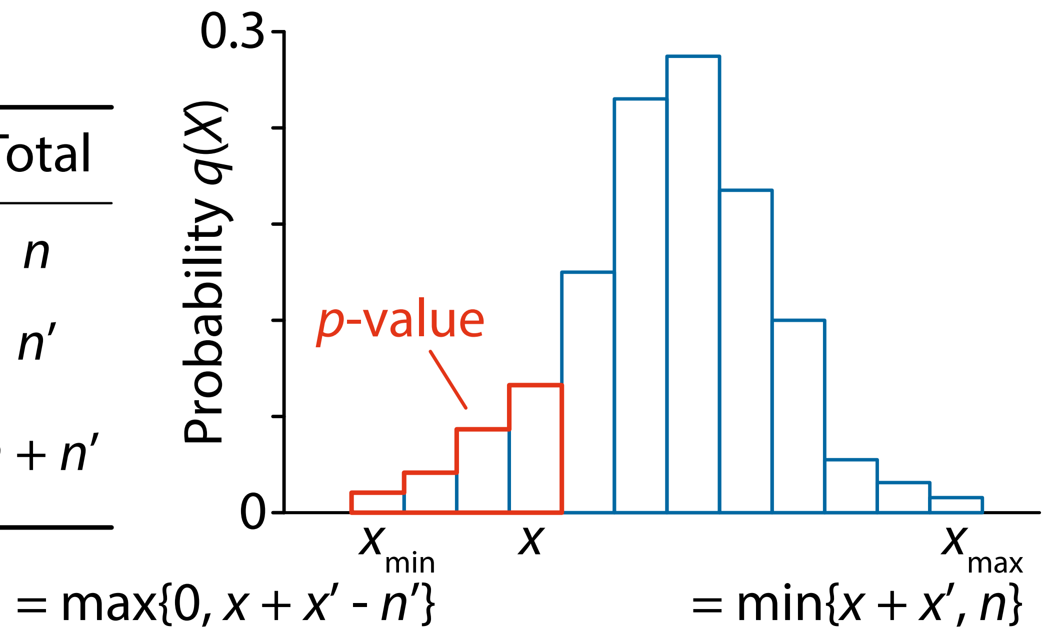


Fisher's Exact Test

- The probability $q(x)$ of obtaining x and x' is given by the hypergeometric distribution:

$$q(x) = \binom{n}{x}\binom{n'}{x'} \Big/ \binom{n+n'}{x+x'}$$

	Occ.	Non-occ.	Total
$\mathscr{X}$	x	$n-x$	n
$\mathscr{X}'$	x'	$n'-x'$	n'
Total	$x+x'$	$(n-x)$ $+ (n'-x')$	$n+n'$



Multiple Testing Correction

- In each test, (p-value $< \alpha$) $\Rightarrow$ statistically significant
- If we test m patterns, αm subgraphs are false positives
  - α: Significance level (predetermined by the user)
- Example in itemset mining:
  - There are 100000 items
  - Number of combinations are 2^{100000}
  - Set significance level $\alpha = 0.01$
  - Number of false positives: $0.01 \cdot 2^{100000} = 10^{30101}$
- FWER: Probability of having more than one false positives among all patterns
  - $\mathrm{FWER} = 1 - (1 - \alpha)^m$ if patterns are independent

Controlling the FWER

- FWER = Pr(FP > 0)
  - FP: Number of false positives
- To achieve FWER = α, change the significance level for each test from α to δ
  - δ: corrected significance level
  - $\delta \leq \alpha$
- Objective is to optimize (maximize) δ:

$$\delta^* = \operatorname*{argmax}_{\delta} \mathrm{FWER}(\delta) \quad \text{s.t. } \mathrm{FWER}(\delta) \leq \alpha$$

  - FWER(δ): FWER at corrected significance level δ
    - Cannot be evaluated in closed form
  - Bonferroni correction is popular: $\delta^*_{\mathrm{Bon}} = \alpha / m$

Westfall-Young Permutation

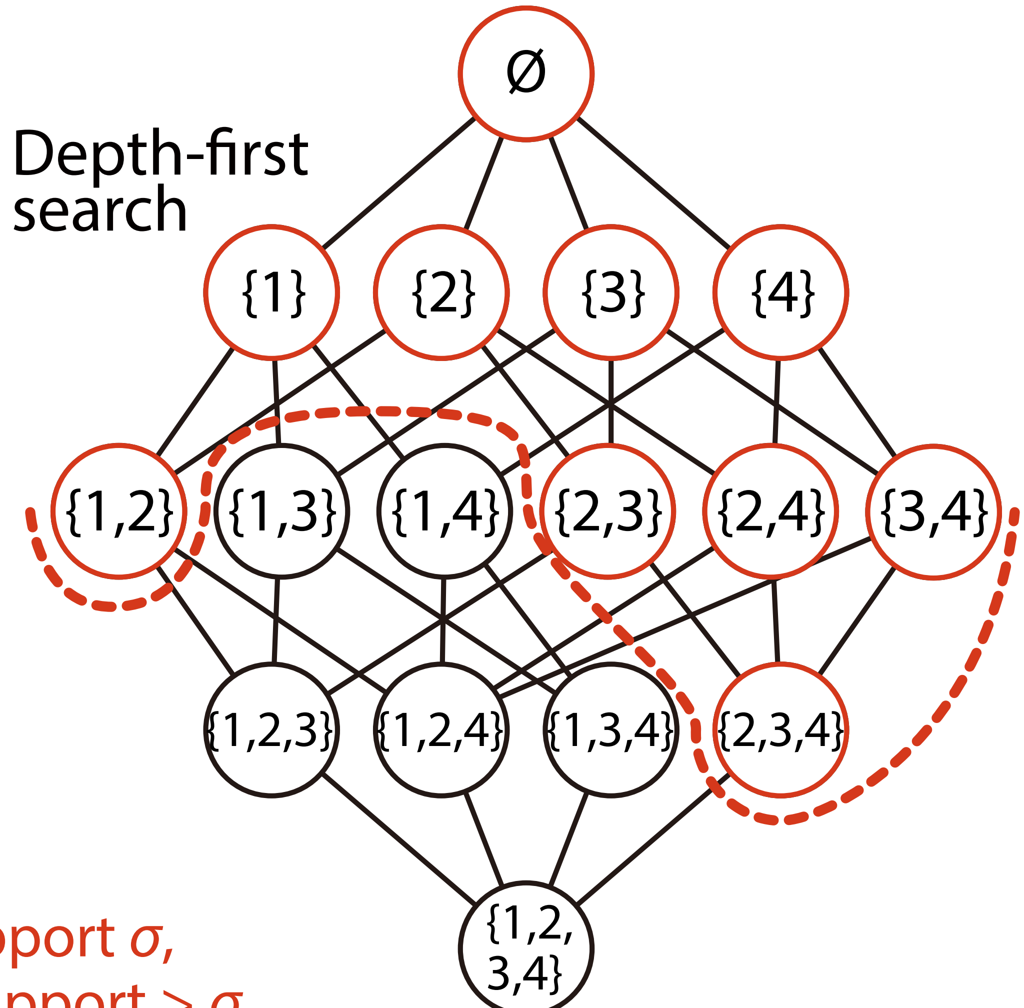
1. Randomly permute class labels
2. Compute p-values for all patterns using the permuted class labels
3. Find the minimum p-value $p_{\min}$ among them
   - $\mathrm{FP} > 0 \iff p_{\min} < \delta$
     - FP: Number of false positives
4. Repeat steps 1 to 3 h times and obtain $p^1_{\min}, p^2_{\min}, \ldots, p^h_{\min}$
   - $\mathrm{FWER}(\delta) \approx |\{\, i : p^i_{\min} \le \delta \,\}| \,/\, h$
5. δ^* is the α-quantile of $p^1_{\min}, p^2_{\min}, \ldots, p^h_{\min}$

Pattern Mining

Transaction data

ID	1 2 3 4
1	1 1 1 1
2	1 1 0 0
3	0 1 0 1
4	0 1 1 1

Task:
Find all patterns
(sets of features)
whose support ≥ 2



Apriori principle:
For a pattern H with support σ,
none of its superset's support > σ

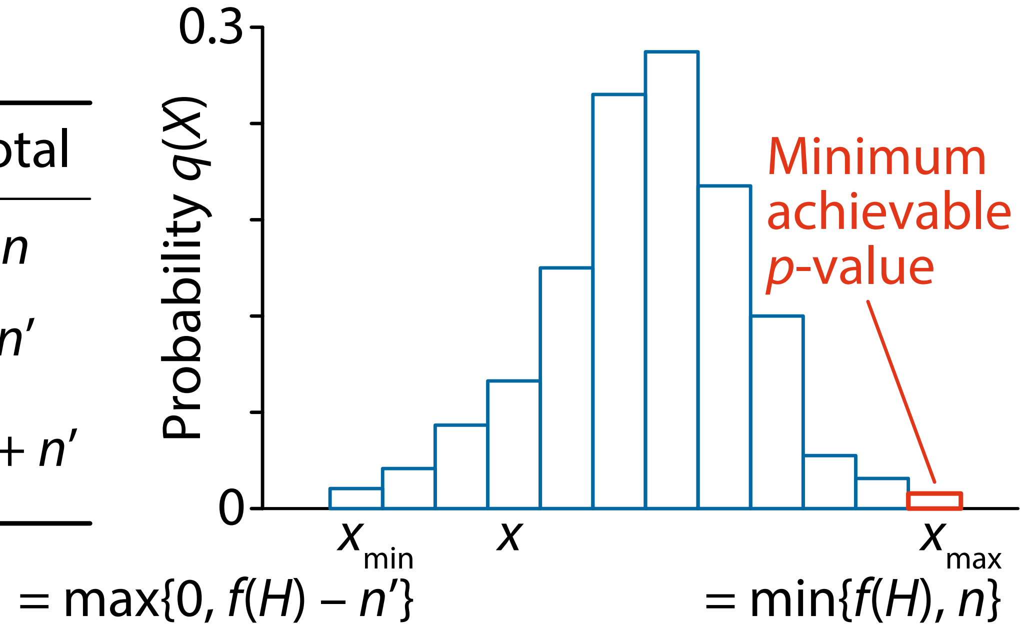
"Westfall-Young light"

- Precompute h permuted labels; $\sigma \leftarrow 1$; $p^i_{\mathrm{min}} \leftarrow 1$
- Westfall-Young light does the following whenever a miner (like LCM) finds a new frequent pattern H:
  - **for** $i \leftarrow 1$ **to** h **do**:
    - $p^i \leftarrow$ the p-value of H for ith permutation
    - $p^i_{\mathrm{min}} \leftarrow \min\{p^i_{\mathrm{min}}, p^i\}$
  - FWER $\leftarrow |\{\, i : p^i_{\mathrm{min}} \leq \Psi(\sigma) \,\}| \,/\, h$
    // $\Psi(\sigma)$ is the min. achievable p-value at σ
  - **while** FWER $> \alpha$ **do**:
    - $\sigma \leftarrow \sigma + 1$ // σ is the minimum support
    - FWER $\leftarrow |\{\, i : p^i_{\mathrm{min}} \leq \Psi(\sigma) \,\}| \,/\, h$
  - Go children of H

Minimum Achievable p-value

- $\Psi(\sigma)$ is the minimum achievable p-value of a pattern H when its support $\sigma = |\{ X \in \mathcal{X} \cup \mathcal{X}' \mid H \subseteq X \}|$
- $\Psi(\sigma) = \min\{ p(x) \mid x_{\min} \le x \le x_{\max} \}$
  - $x_{\min} = \max\{0, \sigma - n'\}$, $x_{\max} = \min\{\sigma, n\}$

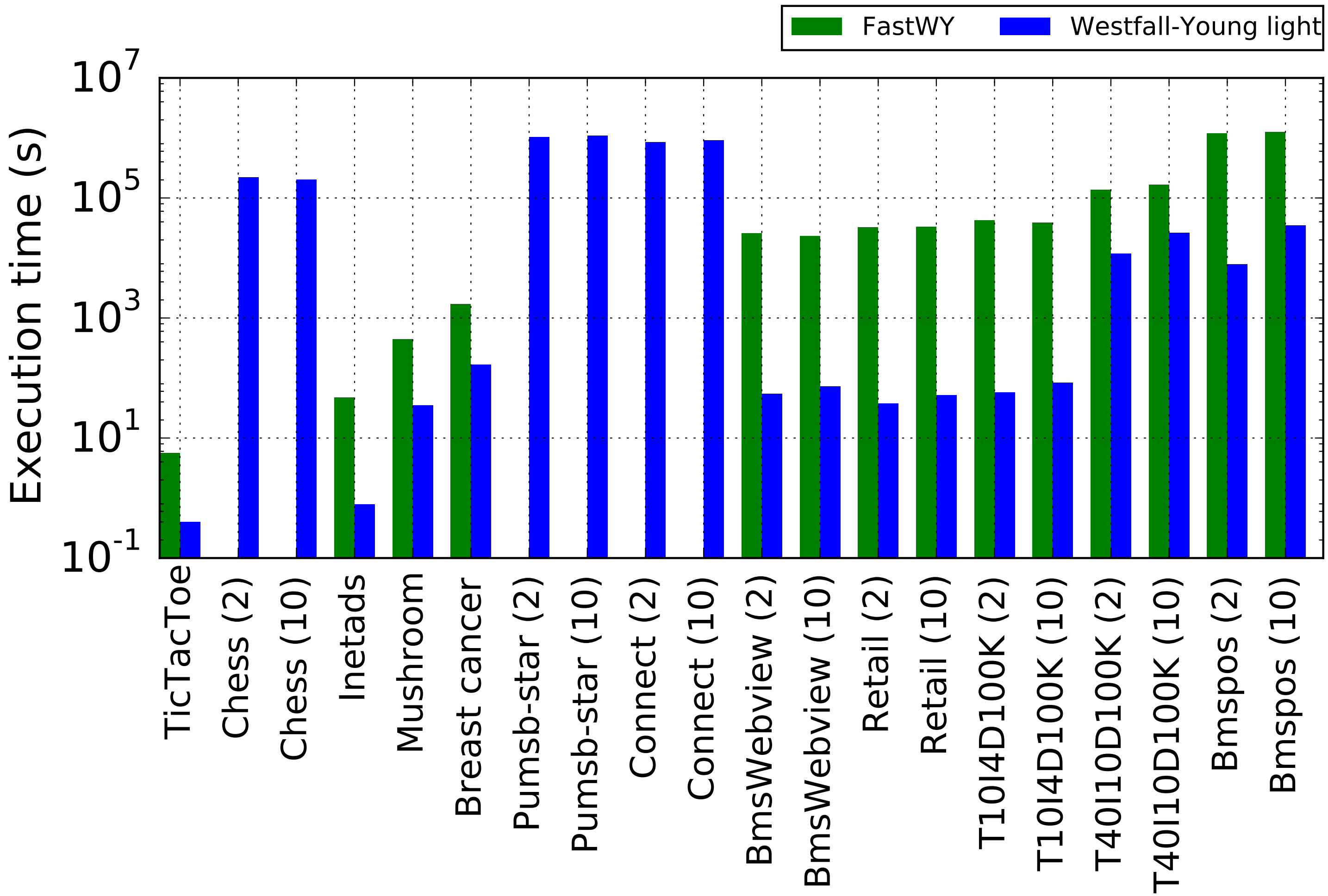
	Occ.	Non-occ.	Total
$\mathcal{X}$	x	$n - x$	n
$\mathcal{X}'$	x'	$n' - x'$	n'
Total	σ	$(n - x) + (n' - x')$	$n + n'$



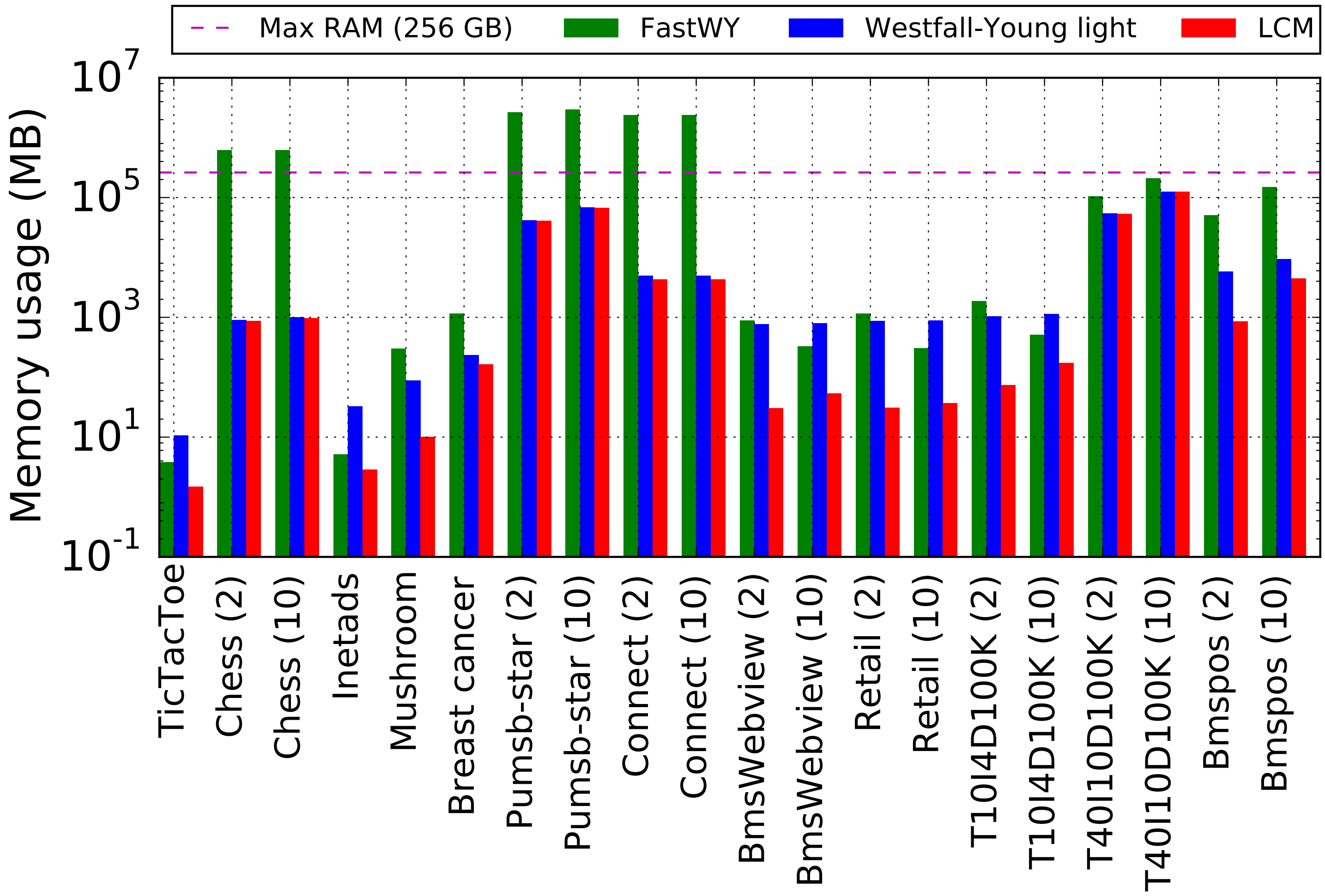
Experiments

- Compare runtime and memory usage of FastWY and Westfall-Young light
  - We reimplemented FastWY in C (x1000 speedup, x10 less memory compared to the Python version)
- Datasets:
  - 20 itemset mining datasets (LCM v3 used as a miner)
  - 12 graph mining datasets (Gaston used as a miner)
- All experiments run on a single 2.5 GHz Intel Xeon CPU with 256 GB of memory

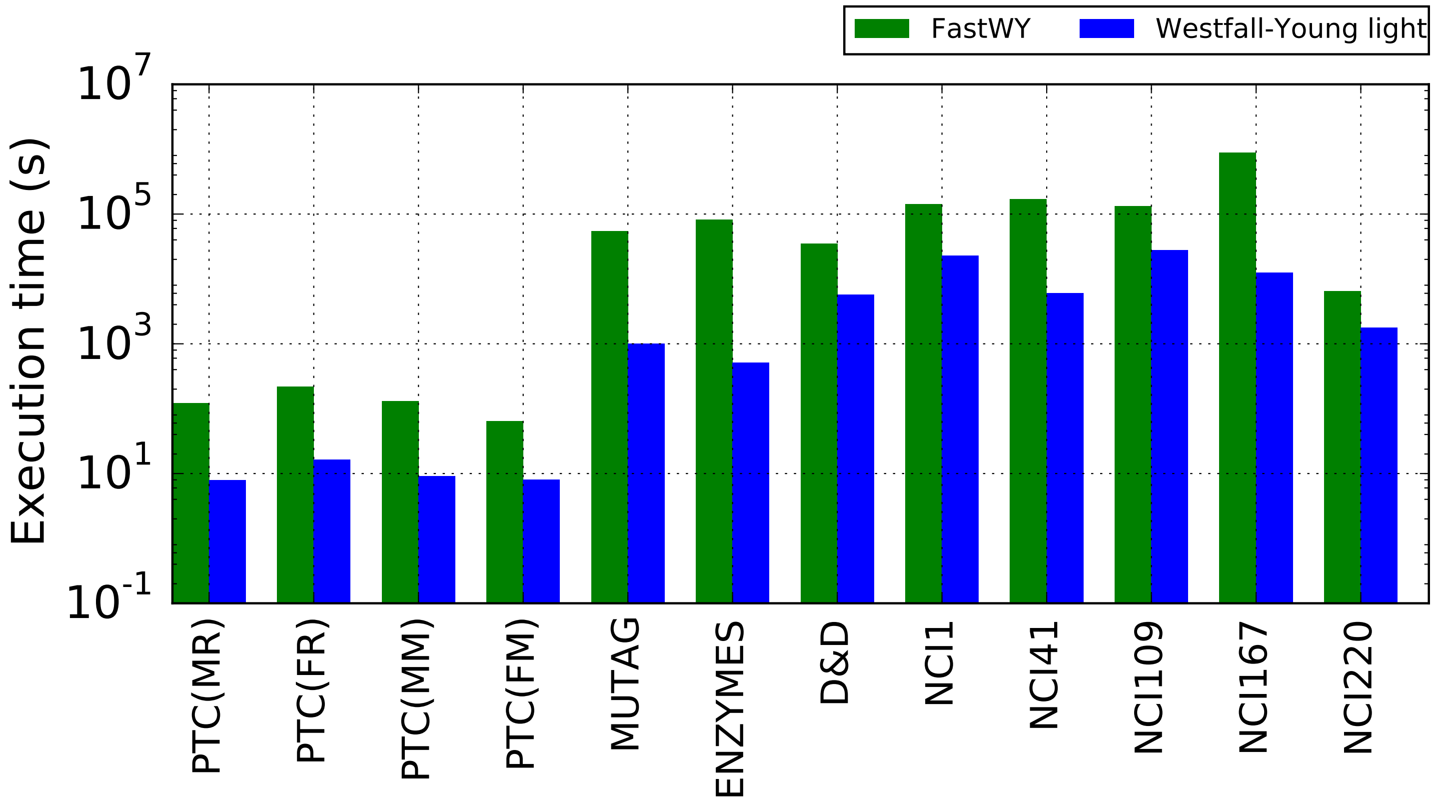
Runtime in Itemset Mining



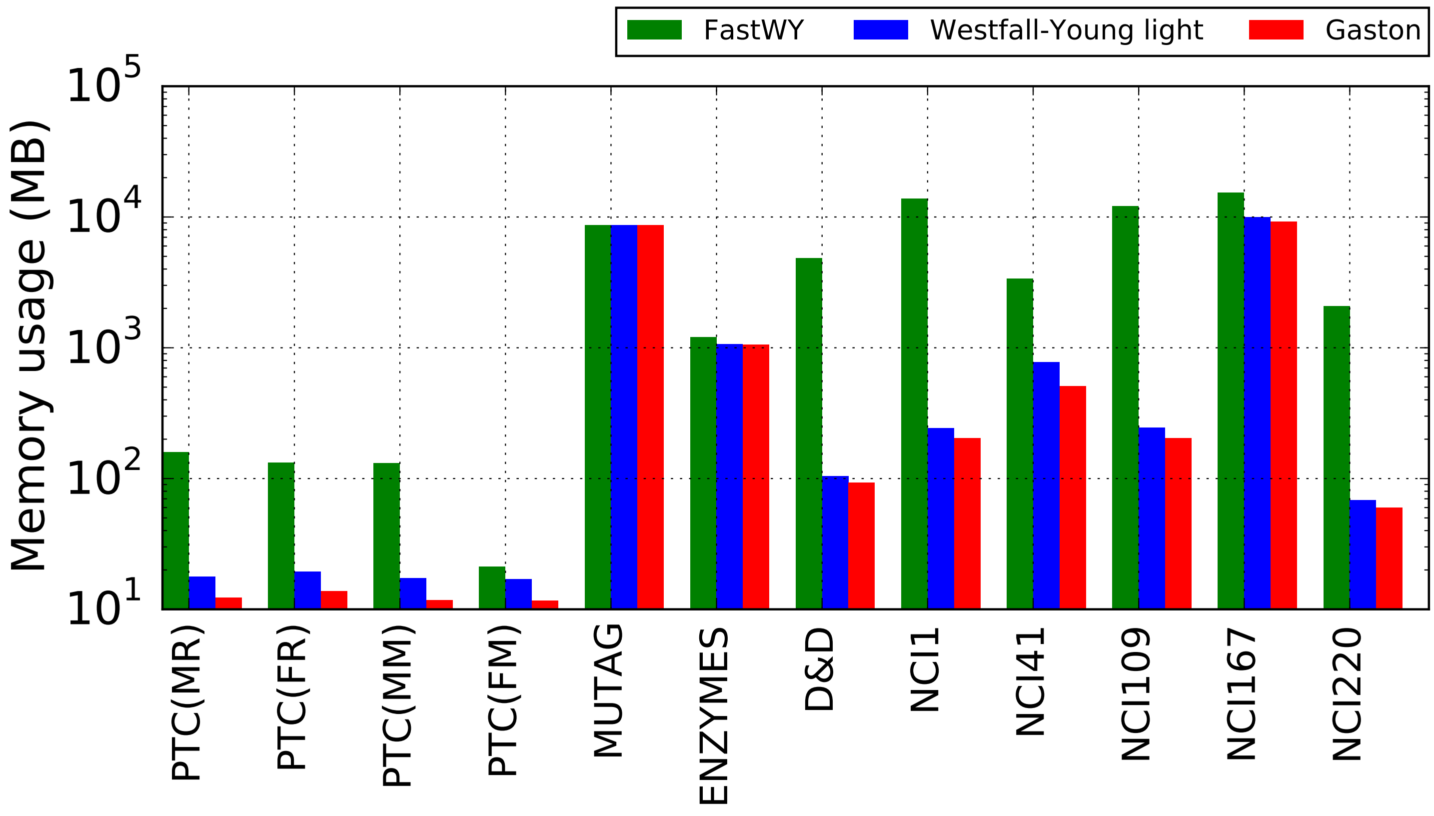
Peak Memory Usage in Itemset Mining



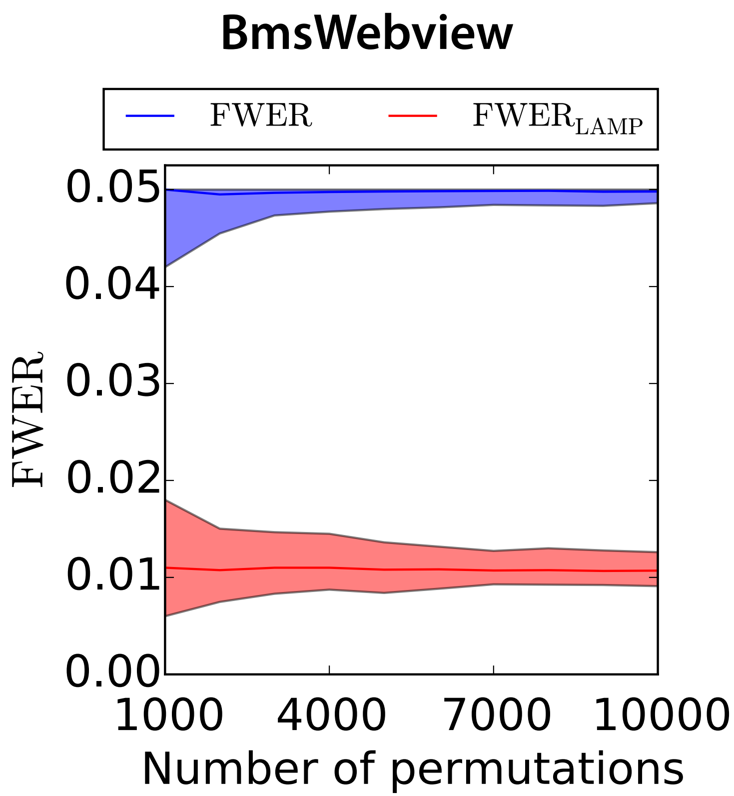
Runtime in Subgraph Mining

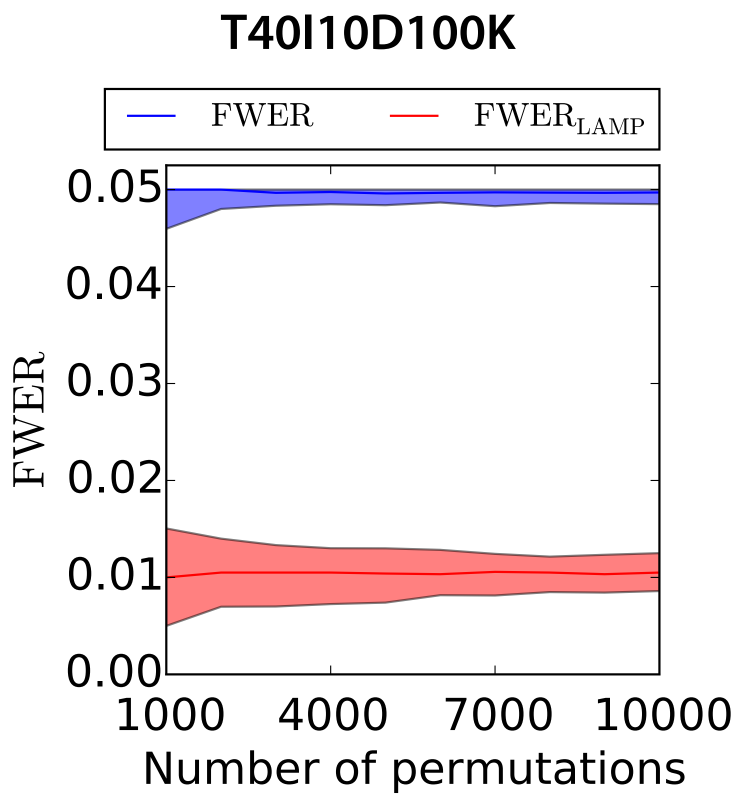


Peak Memory in Subgraph Mining

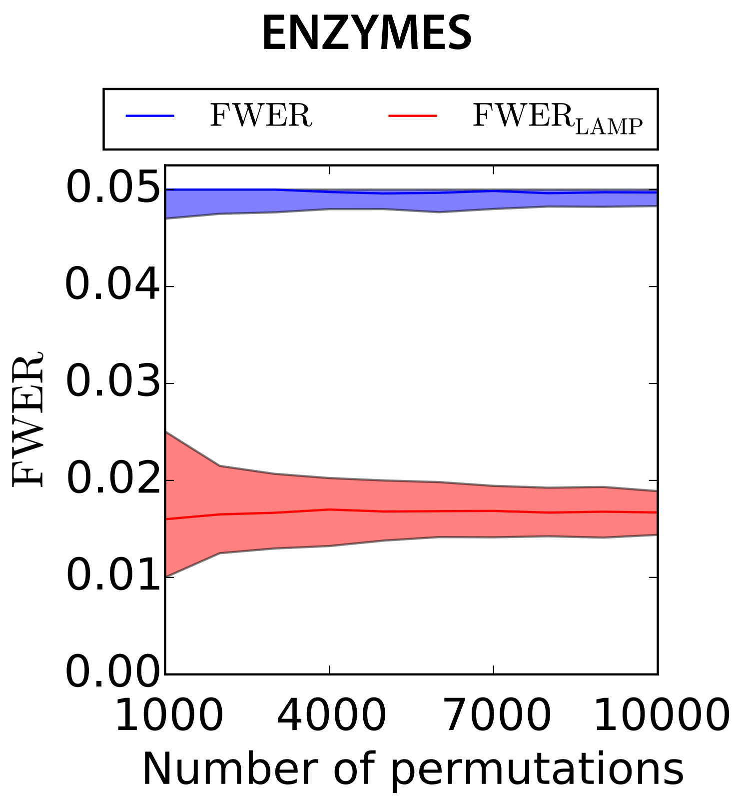


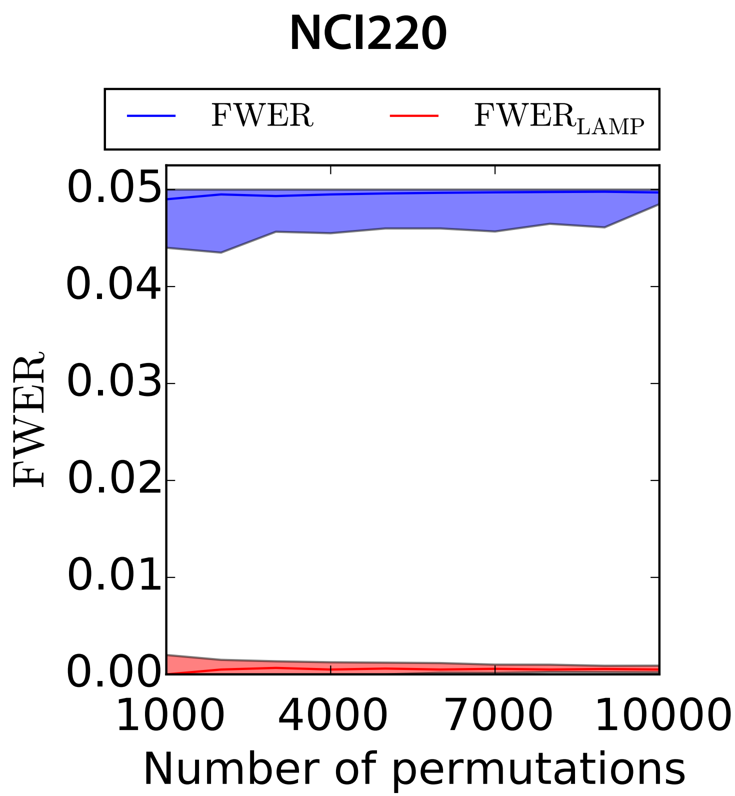
FWER in Itemset Mining





FWER in Subgraph Mining





Conclusion

- Westfall-Young light
  - Code: `http://www.bsse.ethz.ch/mlcb/research/machine-learning/wylight.html`
- The area of significant pattern mining is emerging
  - Find statistically significant combinatorial patterns while controlling false positive rate
- Pattern mining, a classical yet central topic in data mining, can be enriched by introducing statistical assessment
  - Can be applied in scientific fields such as biology