Exploiting discrete test statistics for significant pattern mining

Theory and applications

Felipe Llinares López

D-BSSE, ETH Zürich

February 16th, 2015

D-BSSE
Department of Biosystems
Science and Engineering

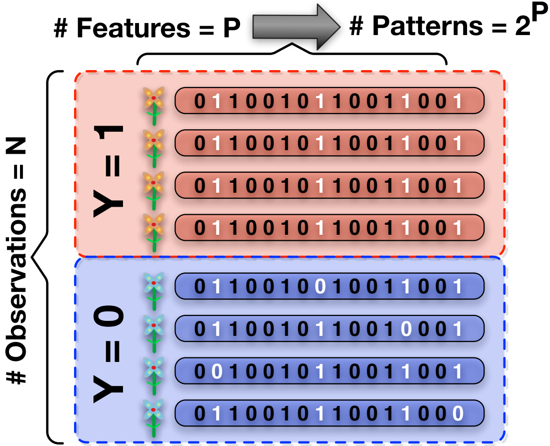
MLCB

Machine
Learning
for
Personalized
Medicine

Outline

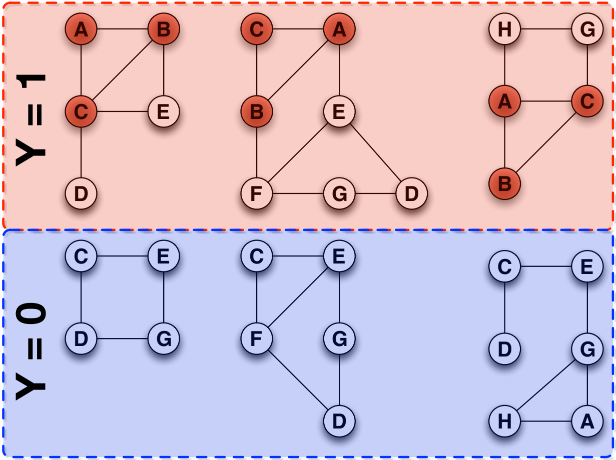
Application: significant itemset mining

- *Example*: Find high-order (multiplicative) combinations of binary predictors associated with class membership



Application: significant subgraph mining

- *Example*: Find molecular motifs statistically associated with drug activity



Problem statement

OUR GOAL

Find **all** patterns whose occurrence within an object is **statistically associated** with class membership, **after correction for multiple testing**

Statistical association testing

Goal

Given $\{(x_i, y_i)\}_{i=1}^{N}$ sampled iid from $p_{\mathbf{XY}}(x, y)$, determine if $\mathbf{X} \not\perp \mathbf{Y}$; i.e. are dependent RVs.

1. Assume $\mathbf{X} \perp \mathbf{Y}$ unless *proven* otherwise
2. Choose a *test statistic* $\mathbf{T}$ to measure the strength of the association between $\mathbf{X}$ and $\mathbf{Y}$ exhibited by the sample $\{(x_i, y_i)\}_{i=1}^{N}$
3. Derive the distribution of $\mathbf{T}$ under the assumption $\mathbf{X} \perp \mathbf{Y}$
4. Given $t = T(\{(x_i, y_i)\}_{i=1}^{N})$, compute the *p-value* as $p = \Pr(\mathbf{T} >= t | \mathbf{X} \perp \mathbf{Y})$
5. *Reject* the assumption $\mathbf{X} \perp \mathbf{Y}$ if $p \leq \alpha$

Statistical association testing in pattern mining

- In pattern mining, every p-value can be obtained from a 2×2 contingency table:

Variables	$\mathbf{X} = 1$	$\mathbf{X} = 0$	Row totals
$\mathbf{Y} = 1$	a	b	n
$\mathbf{Y} = 0$	c	d	$N - n$
Col totals	x	$N - x$	N

- Common test statistics $\mathbf{T}$ for this case are:
  - Fisher's exact test [Fisher, 1922]
  - χ^2*test* [Pearson, 1900]

The multiple hypothesis testing problem

- For a dataset with D patterns $\Rightarrow$ Test $\mathbf{X}_i \perp \mathbf{Y} \; \forall \, i = 1, \ldots, D$
  - We have D different contingency tables
  - Margins n and N the same $\forall \, i = 1, \ldots, D$
  - Margin x depends on $i \Rightarrow x_i$
- When testing the hypothesis $\mathbf{X} \perp \mathbf{Y}$ via $p \leq \alpha$, the probability of a false discovery occurring is α
- What if we test D hypotheses $\mathbf{X}_i \perp \mathbf{Y} \; i = 1, \ldots, D$?
  - *Remark:* $\mathbb{E}[\mathrm{FP}] = \alpha D$
    - $(\alpha = 0.05, D = 2 \cdot 10^5) \Rightarrow 10^4$ false positives on average
    - $(\alpha = 0.05, D = 2 \cdot 10^{10}) \Rightarrow 10^8$ false positives on average
    - $(\alpha = 0.05, D = 2^{2 \cdot 10^5}) \Rightarrow 10^{60205}$ false positives on average

The multiple hypothesis testing problem

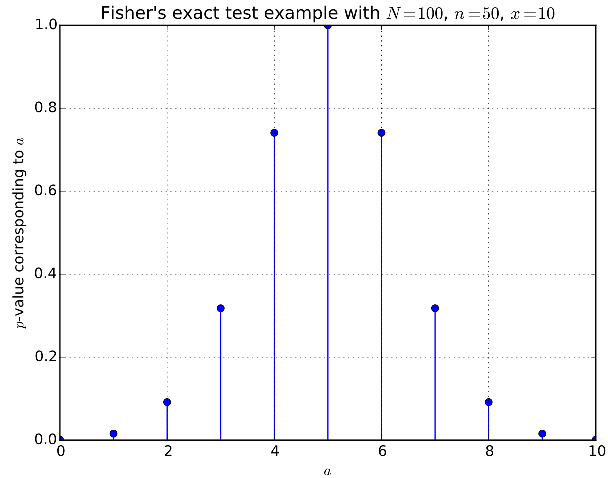
Family-Wise Error Rate (FWER)

The FWER is defined as the probability of producing one or more false discoveries. If one can *guarantee* that $\mathrm{FWER} \leq \alpha$, then the multiple hypothesis testing procedure is said to *control the FWER at level* α

- *Solution*: Reject each hypothesis $\mathbf{X}_i \perp \mathbf{Y}$ $i = 1, \ldots, D$ iff $p_i \leq \delta$, where δ is chosen to ensure $\mathrm{FWER} \leq \alpha$
- *Remark:* $\mathrm{FWER} \ll \alpha$ is not beneficial
- *Bonferroni correction [Bonferroni, 1936]*: Let $\delta = \frac{\alpha}{D}$
- What if, as in pattern mining, D is a gigantic number?

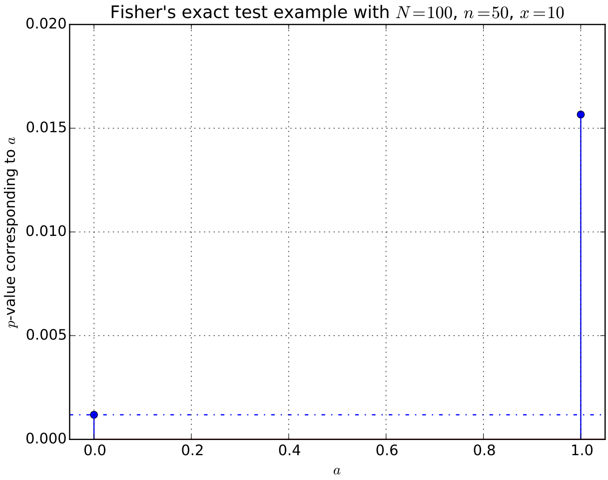
The minimum attainable p-value

In pattern mining, test statistics and attainable p-values are discrete...



The minimum attainable *p*-value

Thus, a minimum attainable *p*-value exists...



The concept of testability

- Given x, n and N, let $\Psi(x) = \min_a p(a, x, n, N)$ be the minimum p-value attainable by the discrete test
    - Remark: Well-defined since (x,n,N) are assumed fixed!
- For each pattern $i = 1, \ldots, D$ the minimum attainable p-value $\Psi(x_i)$ is a function of the pattern support x_i
- If $\Psi(x_i) > \delta$, the i-th pattern can never be significant at corrected level δ
- Define $\mathcal{I}_T(\delta) = \{\, i \in \{\, 1, \ldots, D \,\} \,|\, \Psi(x_i) \leq \delta \,\}$, the set of *testable hypotheses at level* δ
- What are the implications?

An improved Bonferroni Correction for discrete data

- This phenomenon was first exploited in [Tarone, 1990]
- Tarone showed that $\mathrm{FWER} \leq \delta \left|\mathcal{I}_T(\delta)\right|$
- To ensure $\mathrm{FWER} \leq \alpha$ choose $\delta^*_{\mathrm{tar}} = \max \left\{ \delta | \delta \left|\mathcal{I}_T(\delta)\right| \leq \alpha \right\}$
- Usually, $\delta^*_{\mathrm{tar}} \gg \frac{\alpha}{D}$, leading to greatly increased statistical power
- Computing δ^*_{tar} as proposed by Tarone is unfeasible computationally
- In [Terada et al., 2013a], Terada et. al. link Tarone's method to frequent itemset mining, proposing the Limitless-Arity Multiple Testing Procedure (LAMP)

Westfall-Young Light

Felipe Llinares, Mahito Sugiyama, Laetitia Papaxanthos, Karsten Borgwardt

Addressing the dependence between test statistics

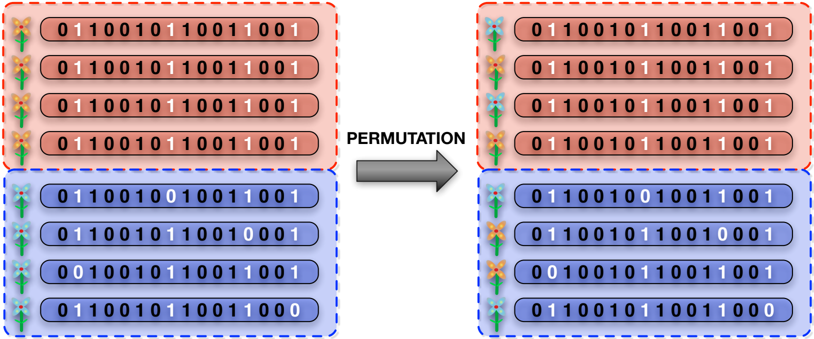
- Tarone's method ignores the dependence structure between patterns: $\mathrm{FWER}(\delta^*_{\mathrm{tar}}) \ll \alpha$ frequently
- The optimal FWER-controlling method would use δ^* such that:

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

- Evaluating $\mathrm{FWER}(\delta)$ in closed-form is not possible
- *Solution:* Use resampling methods, like Westfall-Young (WY) permutation testing [Westfall and Young, 1993]

The Westfall-Young permutation testing procedure

- **Step 1:** Randomly permute class labels $\{y_i\}_i^N$ to obtain $\{\tilde{y}_i\}_i^N$



- By construction, $\mathbf{X}_i \perp \tilde{\mathbf{Y}} \; \forall \, i = 1, \ldots, D$ (i.e. **no** pattern is associated with the class labels)

The Westfall-Young permutation testing procedure

- **Step 2:** Compute the p-values $\tilde{p}_i$ for each pattern $i = 1, \ldots, D$ using the permuted labels $\tilde{\mathbf{Y}}$
  - Since $\mathbf{X}_i \perp \tilde{\mathbf{Y}} \; \forall \, i = 1, \ldots, D$, any pattern for which $\tilde{p}_i \leq \delta$ would be a false positive at level δ
- **Step 3:** Compute $p_{\mathrm{min}} = \min\limits_{i=1,\ldots,D} \tilde{p}_i$
  - $\mathrm{FP} > 0 \Leftrightarrow p_{\mathrm{min}} \leq \delta$
- **Step 4:** Repeat steps 1-3 J times, obtaining $\left\{ p_{\mathrm{min}}^{(j)} \right\}_{j=1}^{J}$
  - $\mathrm{FWER} = \mathrm{Pr}(\mathrm{FP} > 0) \approx \frac{1}{J} \sum_{j=1}^{J} \mathbb{1} \left[p_{\mathrm{min}}^{(j)} \leq \delta \right]$
- **Step 5:** δ^* can be found as the α-quantile of $\left\{ p_{\mathrm{min}}^{(j)} \right\}_{j=1}^{J}$

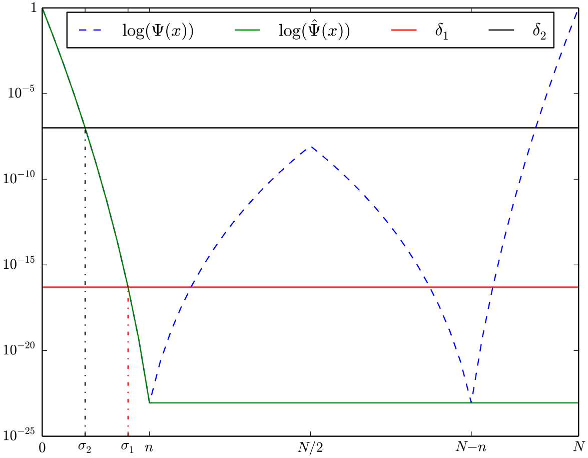
FastWY

- Computing p_{min} naively requires enumerating and computing J p-values for all D patterns
- Terada et. al. propose in [Terada et al., 2013b] the FastWY algorithm as an extension of LAMP to WY permutation testing
- FastWY provides a way to speedup the computation of p_{min} over the naive approach

KEY CONCEPT

The target is computing $p_{\mathrm{min}} = \min\limits_{i=1,\ldots,D} \tilde{p}_i$. If $p'_{\mathrm{min}} = \min\limits_{i \in \mathcal{I}(\delta)} \tilde{p}_i$ satisfies $p'_{\mathrm{min}} \leq \delta$, then $p'_{\mathrm{min}} = p_{\mathrm{min}}$ and the search can be stopped early.

Linking testability and frequent pattern mining



FastWY algorithm

Algorithm 1 FASTWY as proposed in [Terada et al., 2013b]

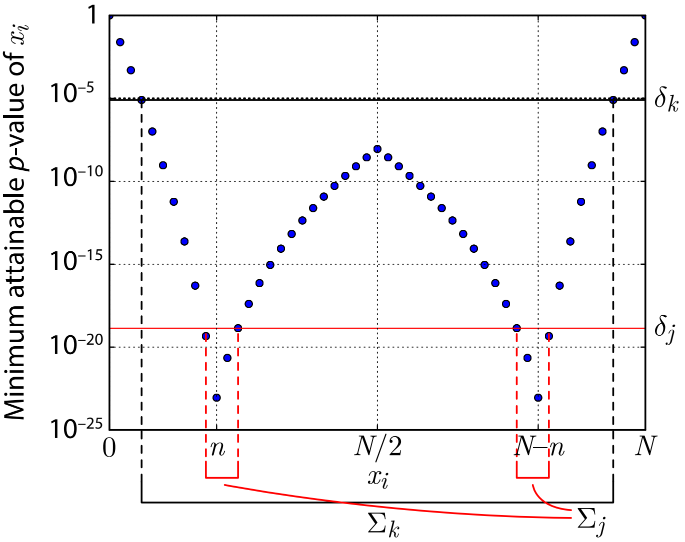
function FASTWY
    for $j = 1, \ldots, J$ **do**
        $\mathbf{y}^{(j)} \leftarrow \mathrm{permute}(\mathbf{y})$
        $\sigma \leftarrow n + 1$
        repeat
            $\sigma \leftarrow \sigma - 1,\ \delta_\sigma \leftarrow \Psi(\sigma)$
            $\hat{\mathcal{I}}_T(\sigma) \leftarrow \mathrm{FPM}(\{1, \ldots, P\}, \sigma)$
            Compute $p_i \ \forall\ i \in \hat{\mathcal{I}}_T(\sigma)$
            $p_{\min}^{(j)} \leftarrow \min_{i \in \hat{\mathcal{I}}_T(\sigma)} p_i$
        until $p_{\min}^{(j)} \leq \delta_\sigma$
    end for
    $\delta^* \leftarrow \alpha$-quantile of $\left\{ p_{\min}^{(j)} \right\}_{j=1}^{J}$
end function

Limitations of FastWY

1. Like LAMP, relies on using a monotonically decreasing surrogate $\hat{\Psi}(x) \leq \Psi(x)$
2. Uses a decremental (in support threshold) search strategy
3. Needs to either repeat pattern mining $J \approx 10^4$ times or store the occurrence list of every frequent pattern
4. Requires computing the whole set $\left\{p_{\mathrm{min}}^{(j)}\right\}_{j=1}^{J}$ exactly
5. As a consequence of (3), with overwhelming probability, some $p_{\mathrm{min}}^{(j)}$ will require mining patterns with very low supports

Westfall-Young light removes all these limitations!

Removing limitation (1)



Removing limitations (1)-(4): the WY-light algorithm

Algorithm 2 WY-light core

function PROCESSPATTERN
  if $x_i \in \Sigma_k$ **then**
    Compute $\tilde{p}_i^{(j)}\ \forall j = 1, \ldots, J$
    $p_{\text{min}}^{(j)} \leftarrow \min(p_{\text{min}}^{(j)}, \tilde{p}_i^{(j)})\ \forall j = 1, \ldots, J$
    $\text{FWER} \leftarrow \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}\left[p_{\text{min}}^{(j)} \leq \delta\right]$
    while $\text{FWER} > \alpha$ **do**
      $k \leftarrow k + 1$
      Update δ_k, Σ_k and $\sigma_k = \min\{x | x \in \Sigma_k\}$
    end while
  end if
  Enumerate all patterns $j \in \text{Children}(i) | x_j \geq \sigma_k$
end function

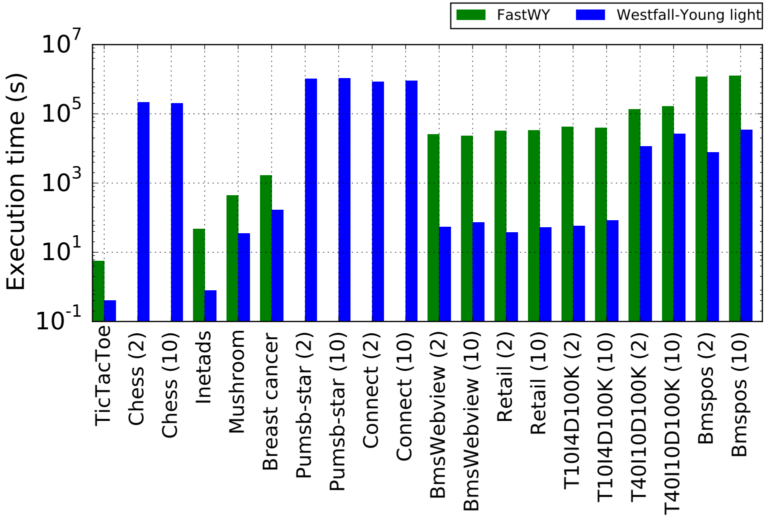
Removing limitations (1)-(4): the WY-light algorithm

Property 1 Whenever a new pattern i is processed, the updated empirical FWER estimate can never decrease

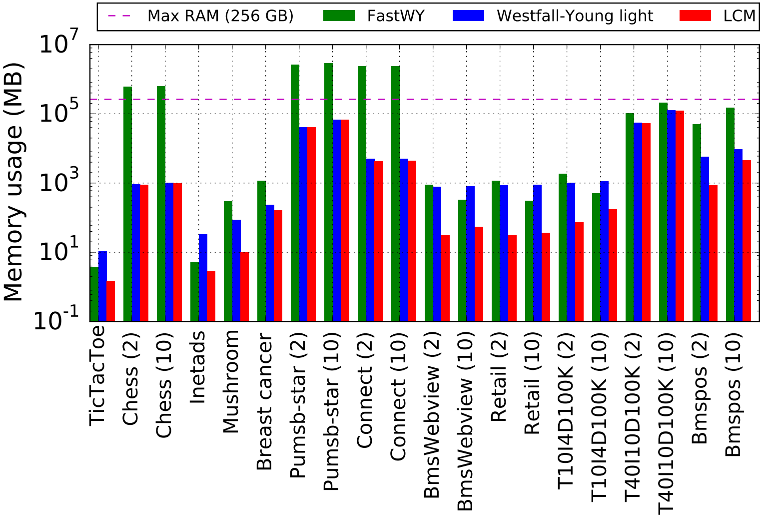
Property 2 $\mathrm{FWER}(\delta)$ for all $\delta \in [0, \delta_k]$ can be evaluated exactly using only the p-values of patterns in $\mathcal{I}_T(\Sigma_k)$.

Property 3 For fixed x_i, n, and N, the computational complexity of evaluating Fisher's exact test p-value $p_i(\gamma)$ for a single value of γ or for all possible values of γ in $[a_{i,\min}, a_{i,\max}]$ is the same and equal to $O(\min\{x_i, n\})$

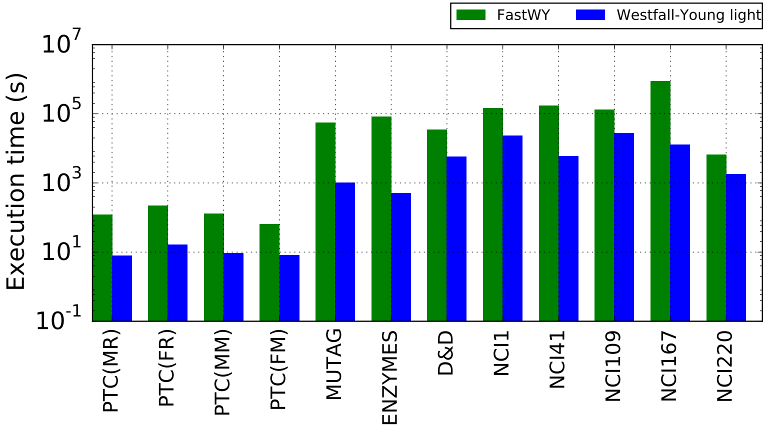
Runtime in itemset mining



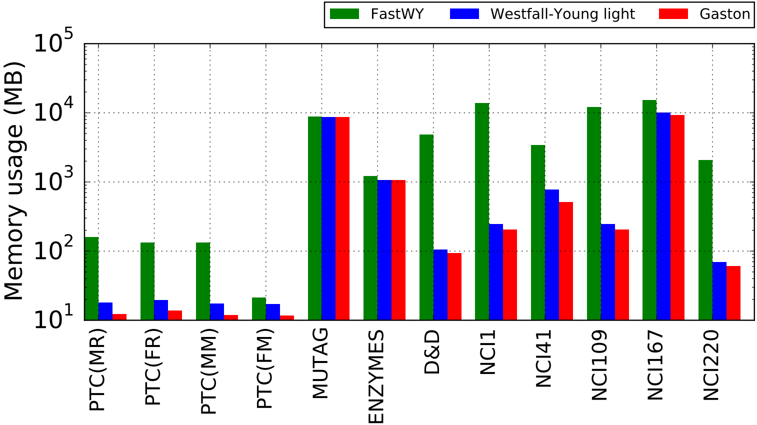
Memory usage in itemset mining



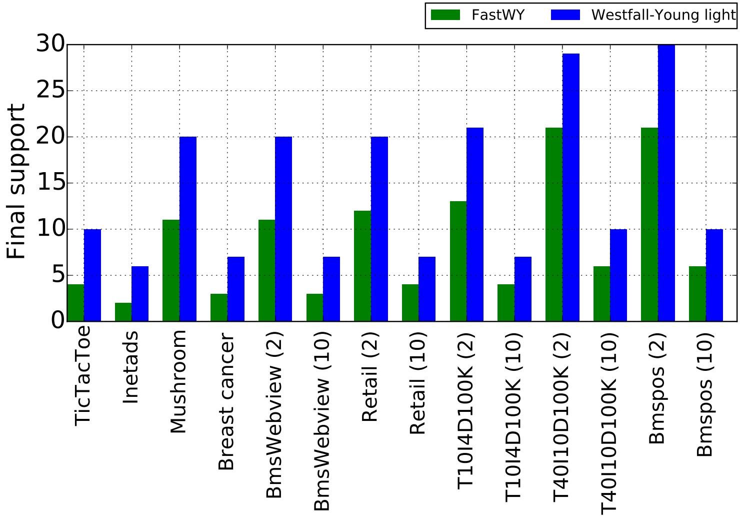
Runtime in subgraph mining



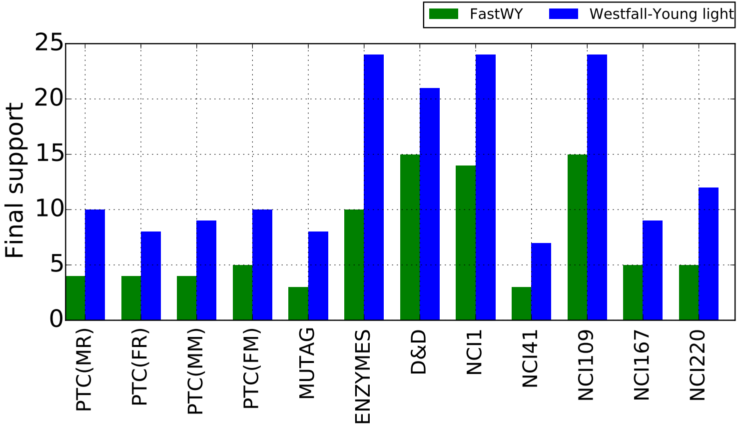
Memory usage in subgraph mining



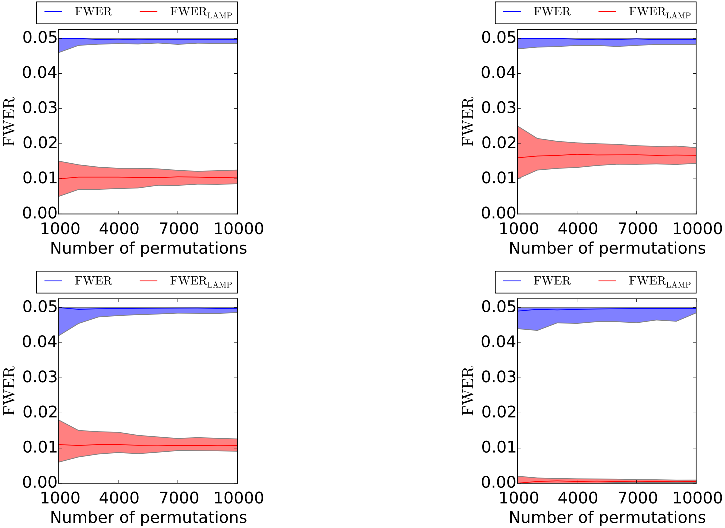
Final support in itemset mining



Final support in subgraph mining



Power comparison: LAMP vs WY



Conclusions

- Discovering patterns significantly associated with class membership is a fundamental problem in data mining
- Rigorous correction for multiple hypothesis testing is mandatory if statistically reliable results are needed
- The discrete nature of test statistics in pattern mining can be exploited to get great gains in statistical power
- Westfall-Young light allows applying the Westfall-Young permutation testing procedure to large-scale datasets
    - Scalable pattern mining under optimal FWER-control!

Thank you!

References I

Bonferroni, C. E. (1936).
Teoria statistica delle classi e calcolo delle probabilità.
Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8:3–62.

Fisher, R. A. (1922).
On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P.
Journal of the Royal Statistical Society, 85(1):87–94.

References II

Pearson, K. (1900).
X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.
Philosophical Magazine Series 6, 50:157–175.

Tarone, R. E. (1990).
A modified bonferroni method for discrete data.
Biometrics, 46(2):515–522.

References III

Terada, A., Okada-Hatakeyama, M., Tsuda, K., and Sese, J. (2013a).
Statistical significance of combinatorial regulations.
Proceedings of the National Academy of Sciences, 110(32):12996–13001.

Terada, A., Tsuda, K., and Sese, J. (2013b).
Fast westfall-young permutation procedure for combinatorial regulation discovery.
In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 153–158.

Westfall, P. H. and Young, S. S. (1993).
Resampling-based multiple testing.
Statistics in Medicine, 13(10):1084–1086.

References IV