June 20–23, 2017
MCP 2017

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

PRESTO
SAKIGAKE

Significant Pattern Mining on Graphs

Mahito Sugiyama (NII, PRESTO)

Literature

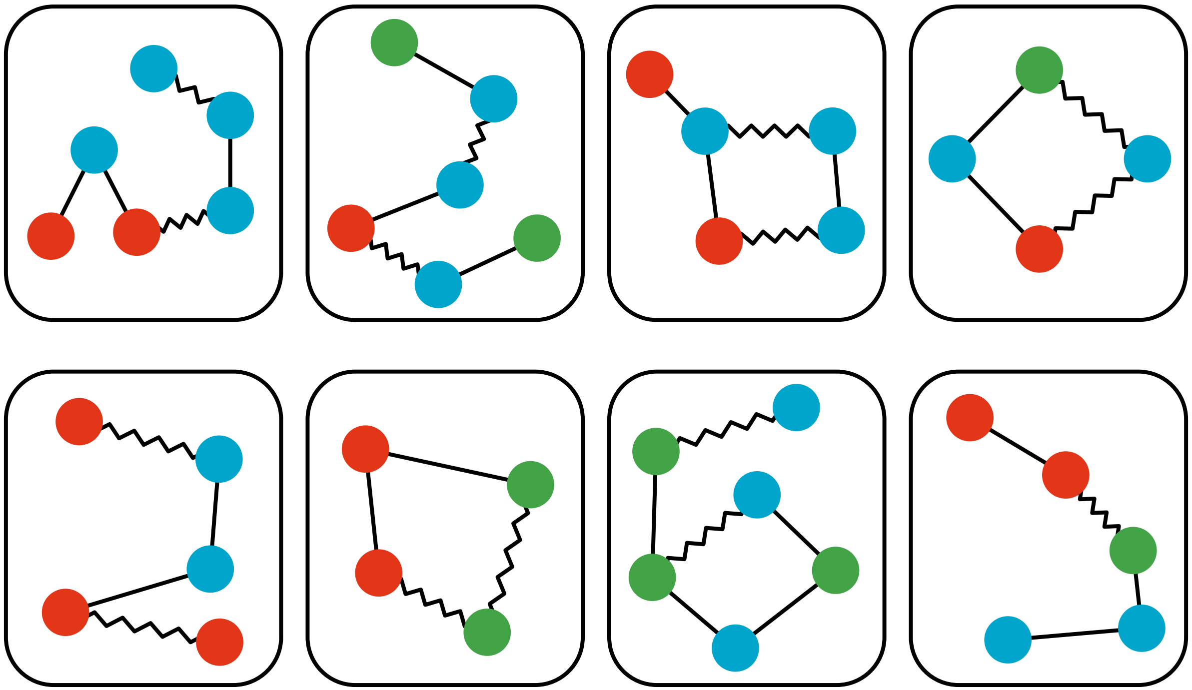
- Sugiyama, M., Llinares-López, F., Kasenburg, N., Borgwardt, K.:
  Significant Subgraph Mining with Multiple Testing Correction,
  SIAM SDM 2015

- Llinares-López, F., Sugiyama, M., Papaxanthos, L., Borgwardt, K.:
  Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing,
  ACM SIGKDD 2015

Subgraph Mining

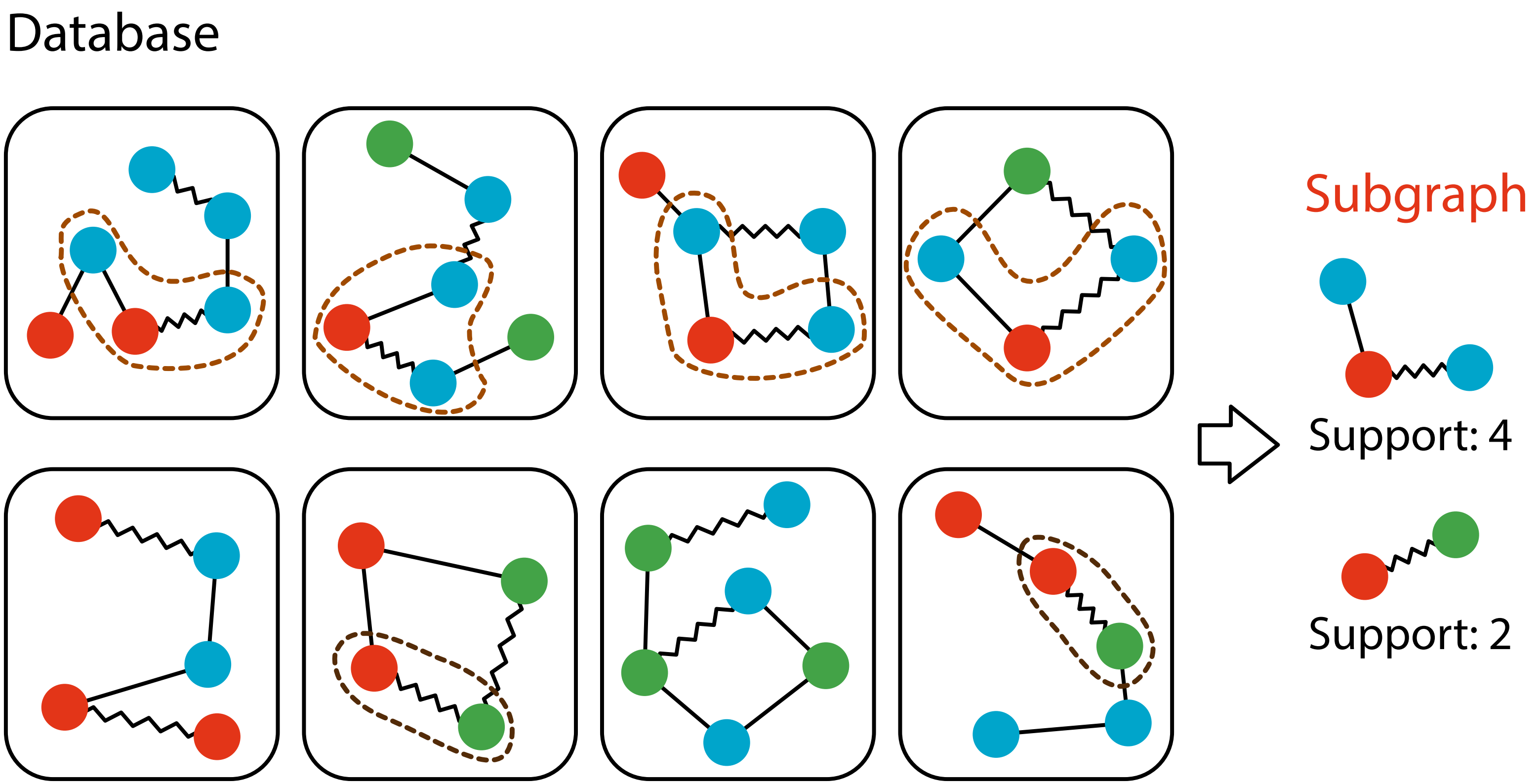
- Find interesting subgraphs from graph databases

Database



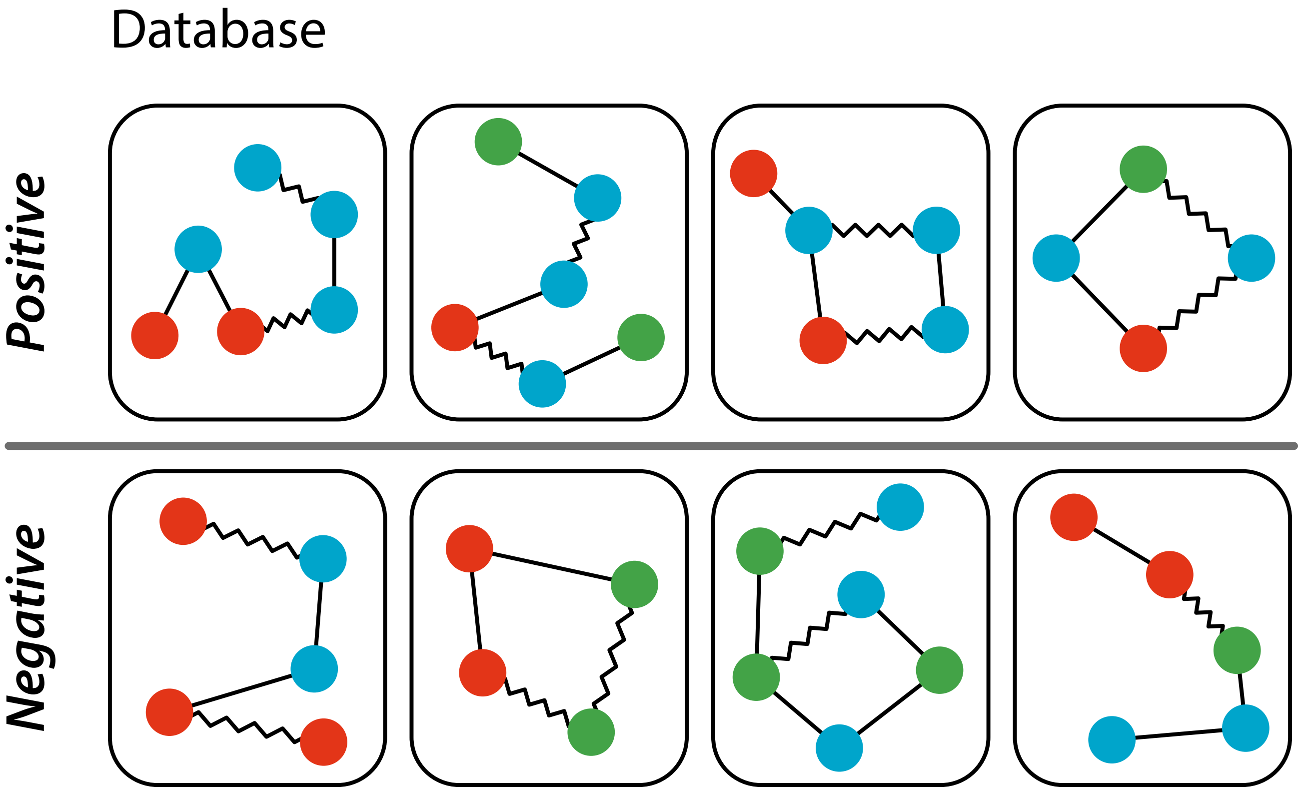
Subgraph Mining

- Find interesting subgraphs from graph databases



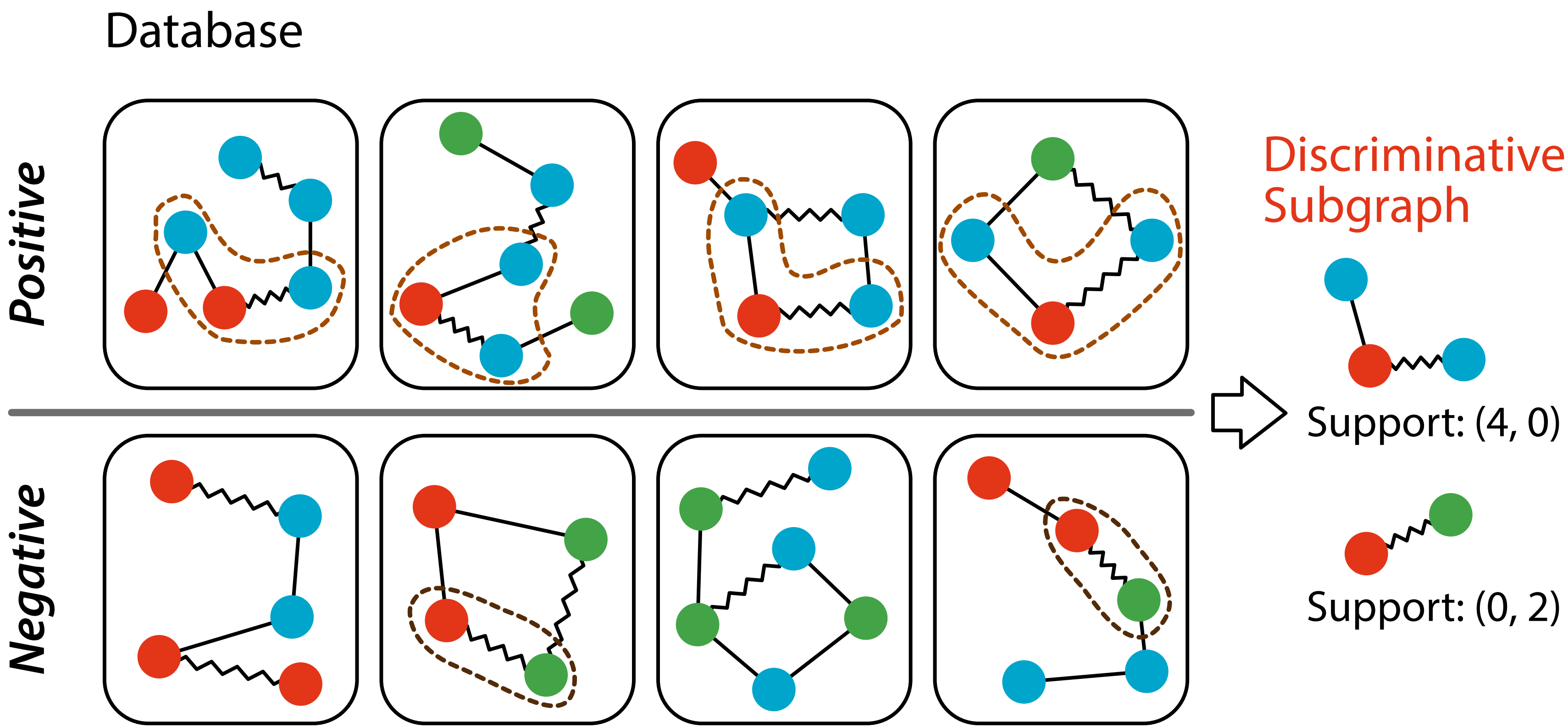
Discriminative Subgraph Mining

- Find discriminative subgraphs from supervised data (e.g. Drug discovery)



Discriminative Subgraph Mining

- Find discriminative subgraphs from supervised data (e.g. Drug discovery)



Challenges and Solutions

- In discriminative subgraph mining:

1. How to measure the discriminability of subgraphs?
2. How to enumerate all discriminative subgraphs?

Challenges and Solutions

- In discriminative subgraph mining:

1. How to measure the discriminability of subgraphs?
2. How to enumerate all discriminative subgraphs?

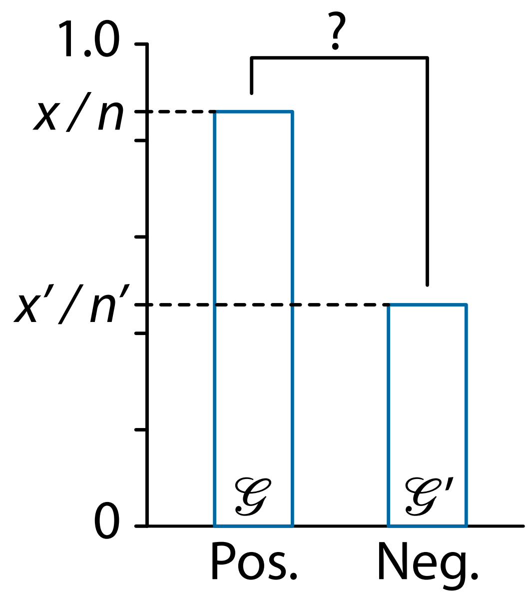
- *Answer to 1:*
  - Compute the p-value via ***statistical hypothesis testing***
  - Discriminative subgraph $\iff$ (Statistically) Significant subgraph
- *Answer to 2:*
  - Integrate evaluation of discriminability and enumeration of subgraphs

Computing p-value of Subgraph

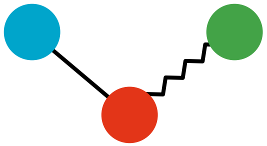
- Given positive and negative sets of graphs $\mathcal{G}, \mathcal{G}'$
  - $|\mathcal{G}| = n, |\mathcal{G}'| = n' \; (n \le n')$
- The p-value of each subgraph H is determined by the Fisher's exact test
  - $x = |\{\, G \in \mathcal{G} \mid H \sqsubseteq G \,\}|$

	Occ.	Non-occ.	Total
$\mathcal{G}$ (Pos.)	x	$n - x$	n
$\mathcal{G}'$ (Neg.)	x'	$n' - x'$	n'
Total	$x + x' = \sigma$	$(n - x) + (n' - x')$	$n + n'$

Support



Multiple Testing



	Occ.	Non-occ.	Total
Positive	4	0	4
Negative	2	2	4
Total	6	2	8

Fisher's exact test: p-value = 0.429

Multiple Testing

	Occ.	Non-occ.	Total
Positive	3	1	4
Negative	1	3	4
Total	4	4	8

Fisher's exact test: *p*-value = 0.486

Pos

Neg

Tot

Fisher's exact test: *p*-value = 0.429

Multiple Testing

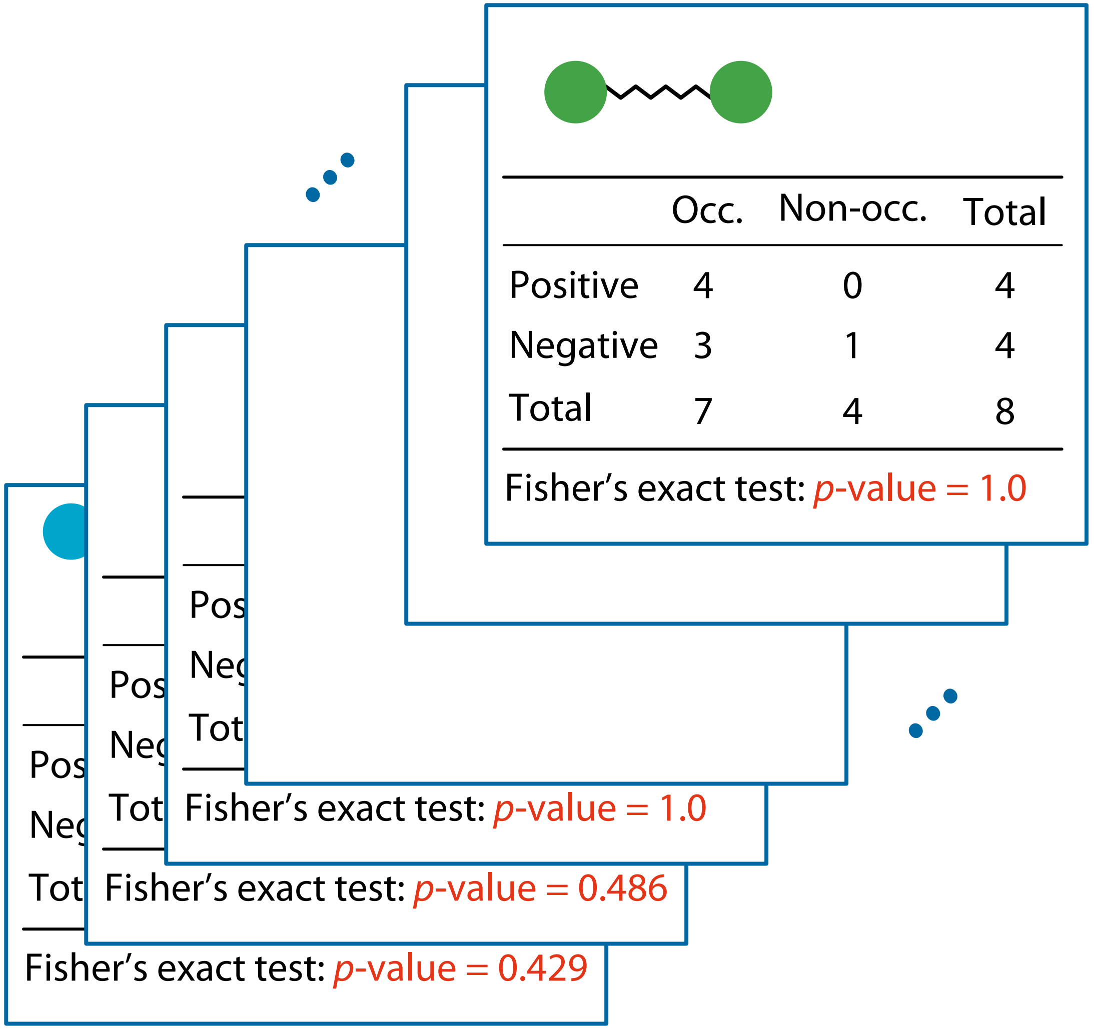
	Occ.	Non-occ.	Total
Positive	2	2	4
Negative	2	2	4
Total	4	4	8

Fisher's exact test: *p*-value = 1.0

Fisher's exact test: *p*-value = 0.486

Fisher's exact test: *p*-value = 0.429

Multiple Testing



Multiple Testing

	Occ.	Non-occ.	Total
Positive	4	0	4
Negative	3	1	4
Total	7	4	8

Fisher's exact test: *p*-value = 1.0

Fisher's exact test: *p*-value = 1.0

Fisher's exact test: *p*-value = 0.486

Fisher's exact test: *p*-value = 0.429

Task: Enumerate all significant subgraphs while controlling the FWER

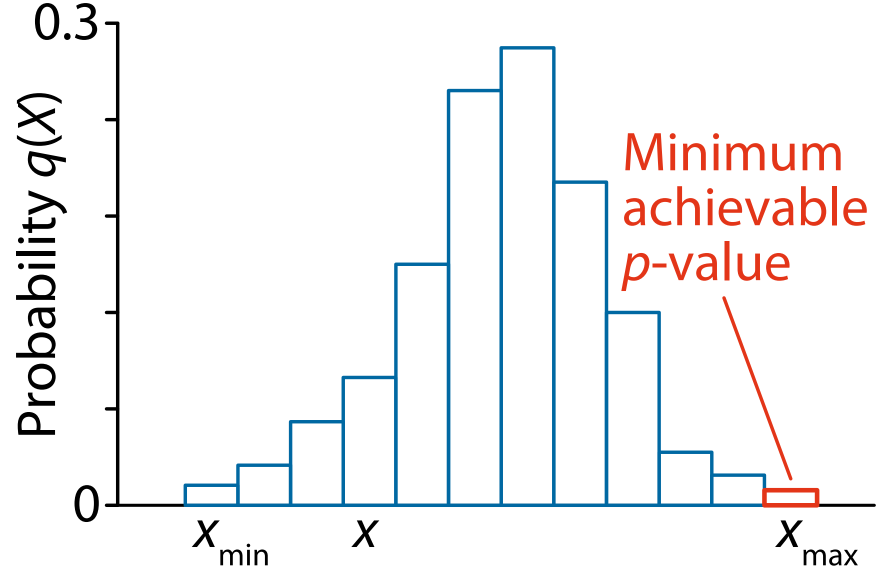
Massive number of (infinitely many) subgraphs (Combinatorial explosion!)

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

- Consider the minimum achievable *p*-value $\Psi(\sigma)$ of a subgraph H for 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
$\mathcal{G}$ (Pos.)	x	$n - x$	n
$\mathcal{G}'$ (Neg.)	x'	$n' - x'$	n'
Total	$x + x' = \sigma$	$(n - x) + (n' - x')$	$n + n'$

Support



$= \max\{0, f(H) - n'\}$ $= \min\{f(H), n\}$

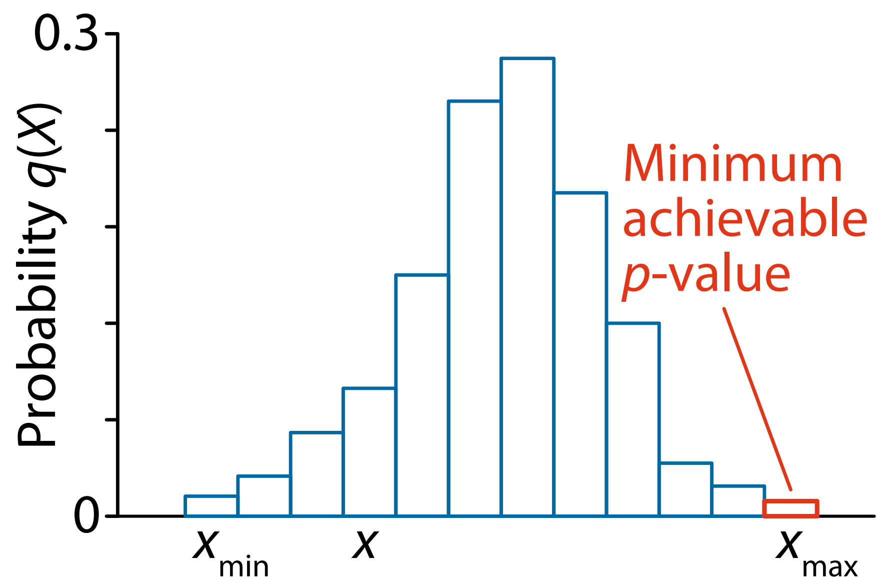
Computing $\Psi(\sigma)$

- Consider the minimum achievable p-value $\Psi(\sigma)$ of a subgraph H for its support $\sigma = |\{ X \in \mathcal{X} \cup \mathcal{X}' \mid H \subseteq X \}|$

$$\Psi(\sigma) = \binom{n}{\sigma} \Big/ \binom{n + n'}{\sigma}$$

	Occ.	Non-occ.	Total
$\mathscr{G}$ (Pos.)	σ	$n - \sigma$	n
$\mathscr{G}'$ (Neg.)	0	n'	n'
Total	σ	$(n - \sigma) + n'$	$n + n'$

Most biased case ($\sigma < n$)



$= \max\{0, f(H) - n'\}$ $= \min\{f(H), n\}$

Testability

- Consider the minimum achievable p-value $\Psi(\sigma)$ of a subgraph H for its support $\sigma = |\{ X \in \mathcal{X} \cup \mathcal{X}' \mid H \subseteq X \}|$

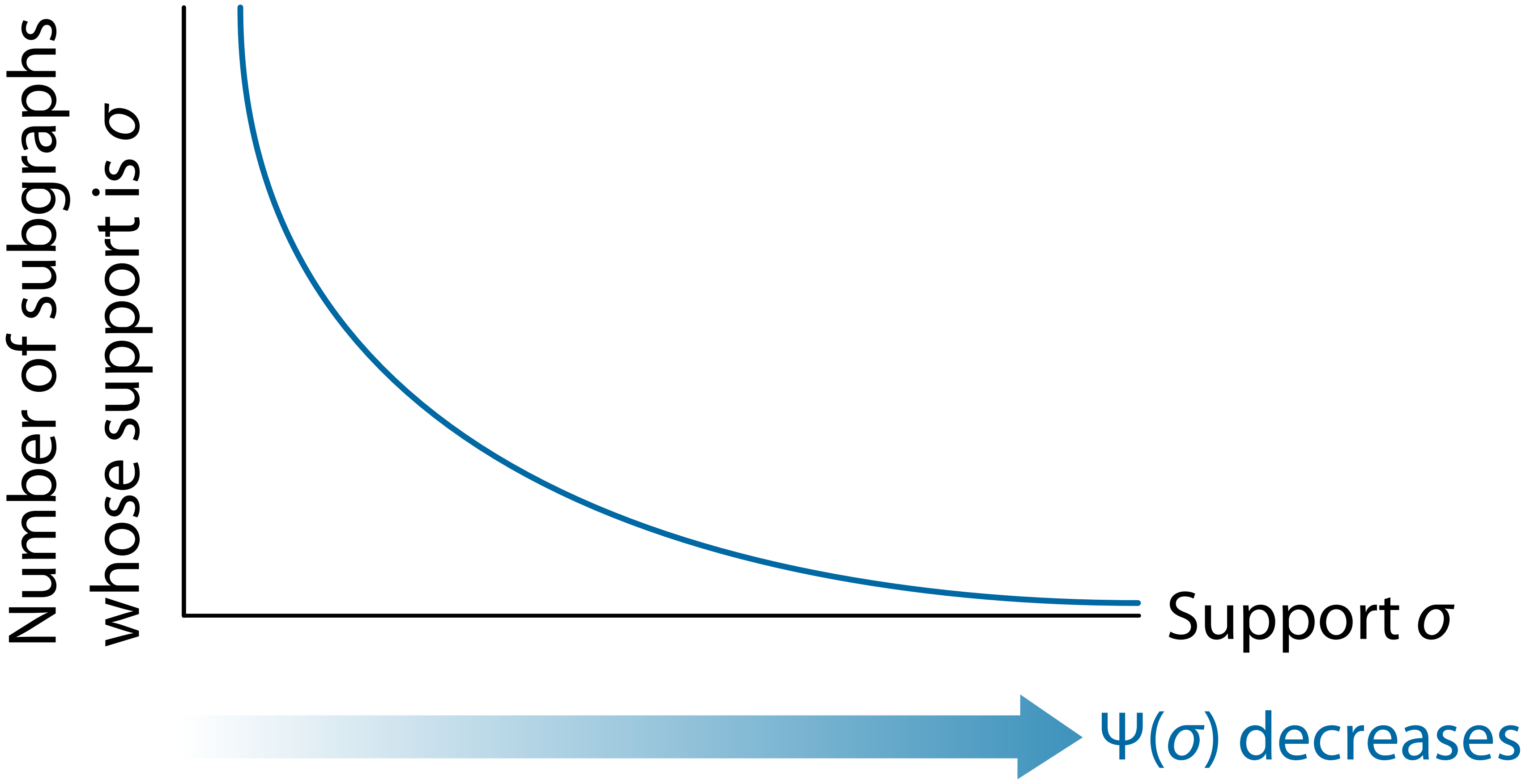
$$\Psi(\sigma) = \binom{n}{\sigma} \bigg/ \binom{n + n'}{\sigma}$$

- Tarone (1990) pointed out (and Terada et al. (2013) revisited):

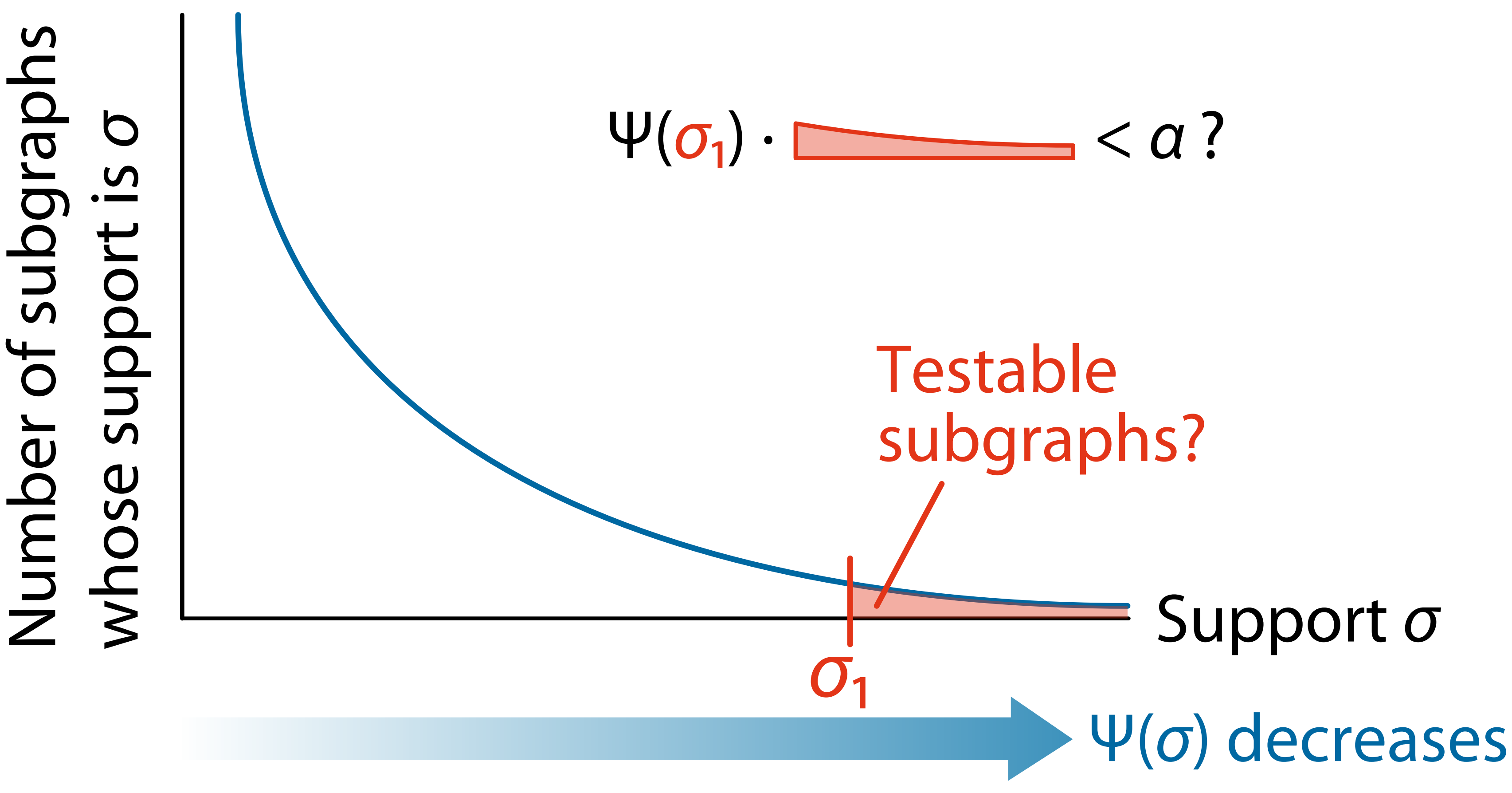
  For a subgraph H with its support σ, if the minimum achievable p-value Ψ(σ) is larger than the significance threshold, this is untestable and we can ignore it

  – Significance threshold = α / [# testable subgraphs]
  – Untestable subgraphs can never be significant

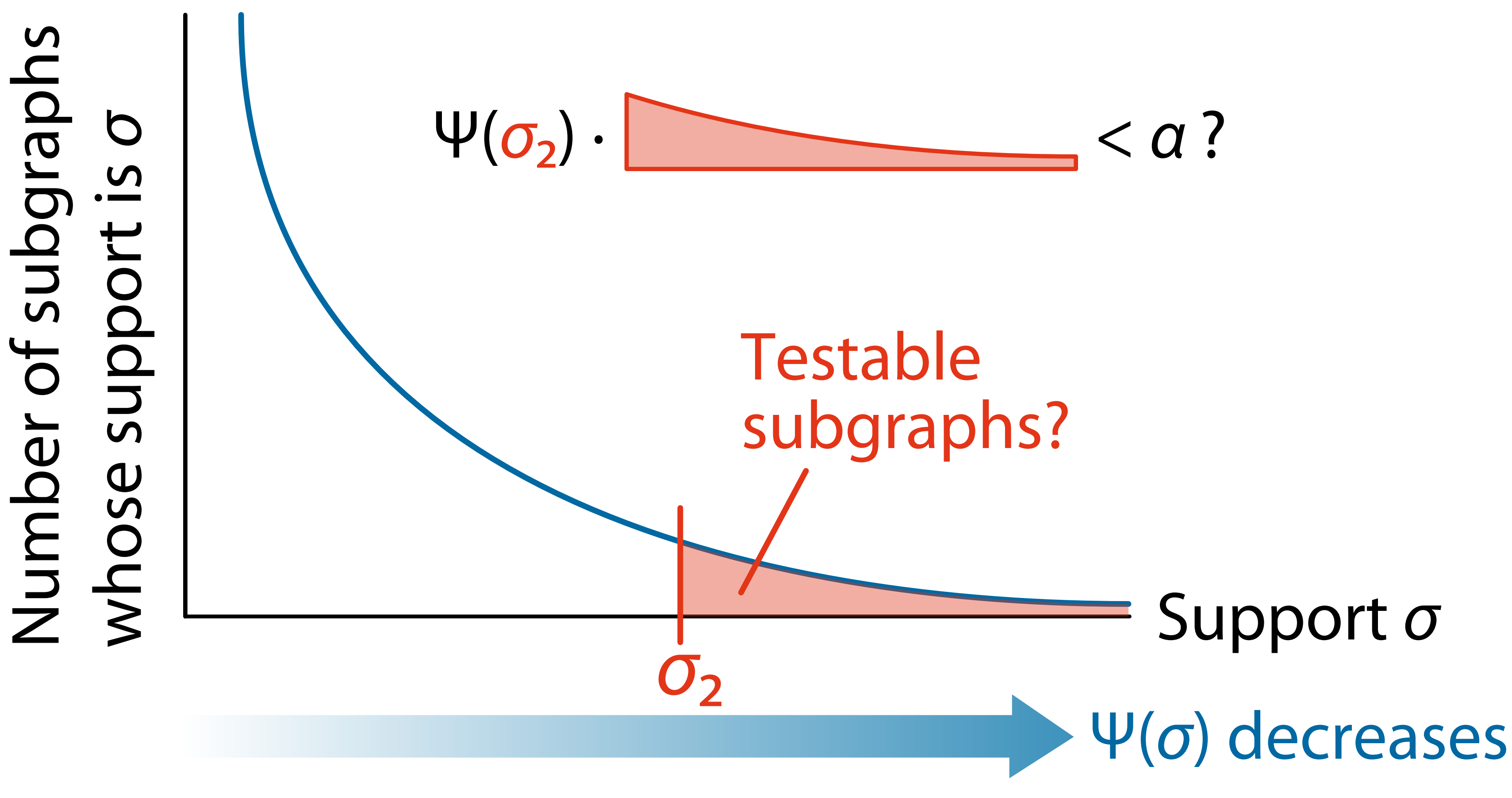
Finding Testable Subgraphs



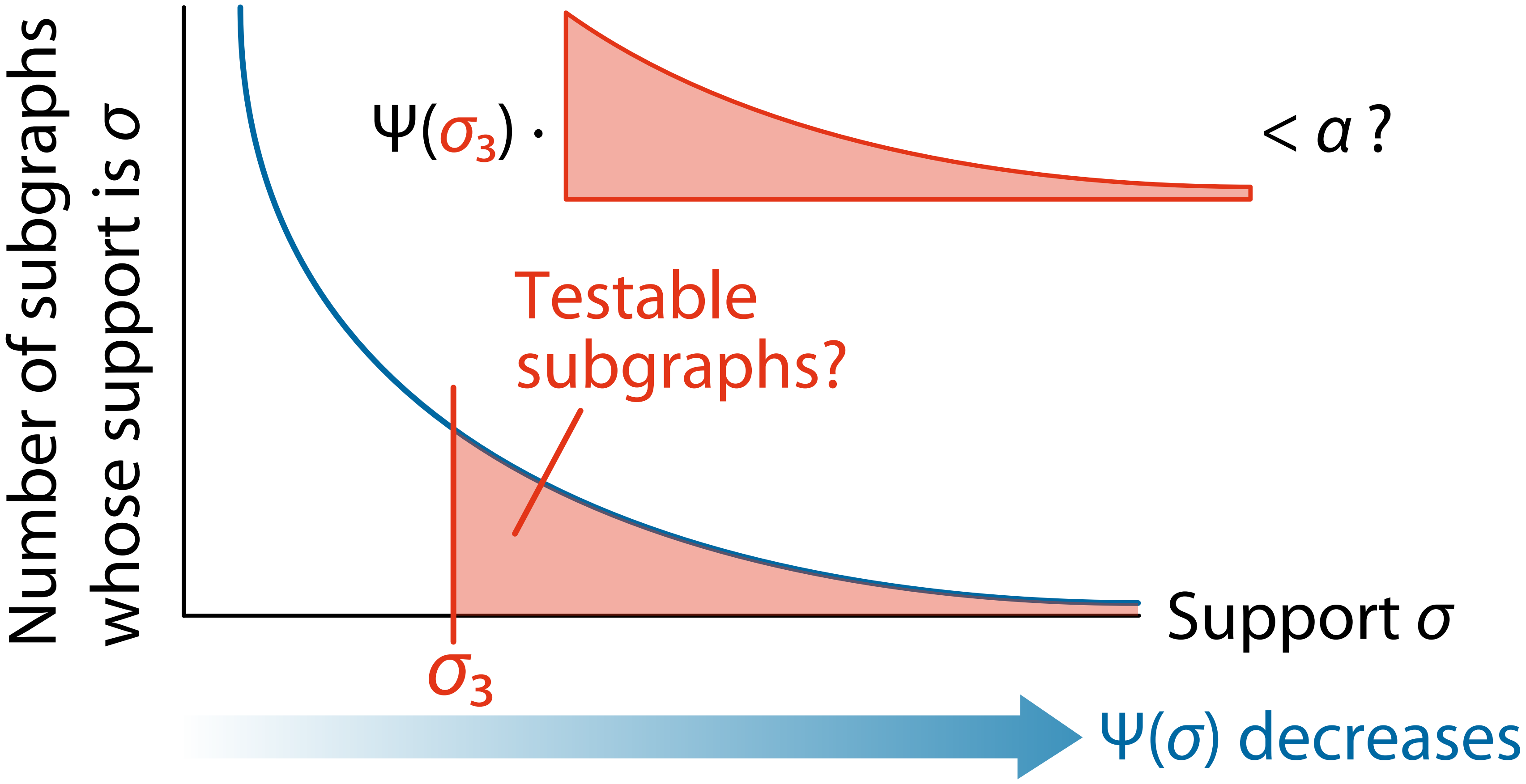
Finding Testable Subgraphs



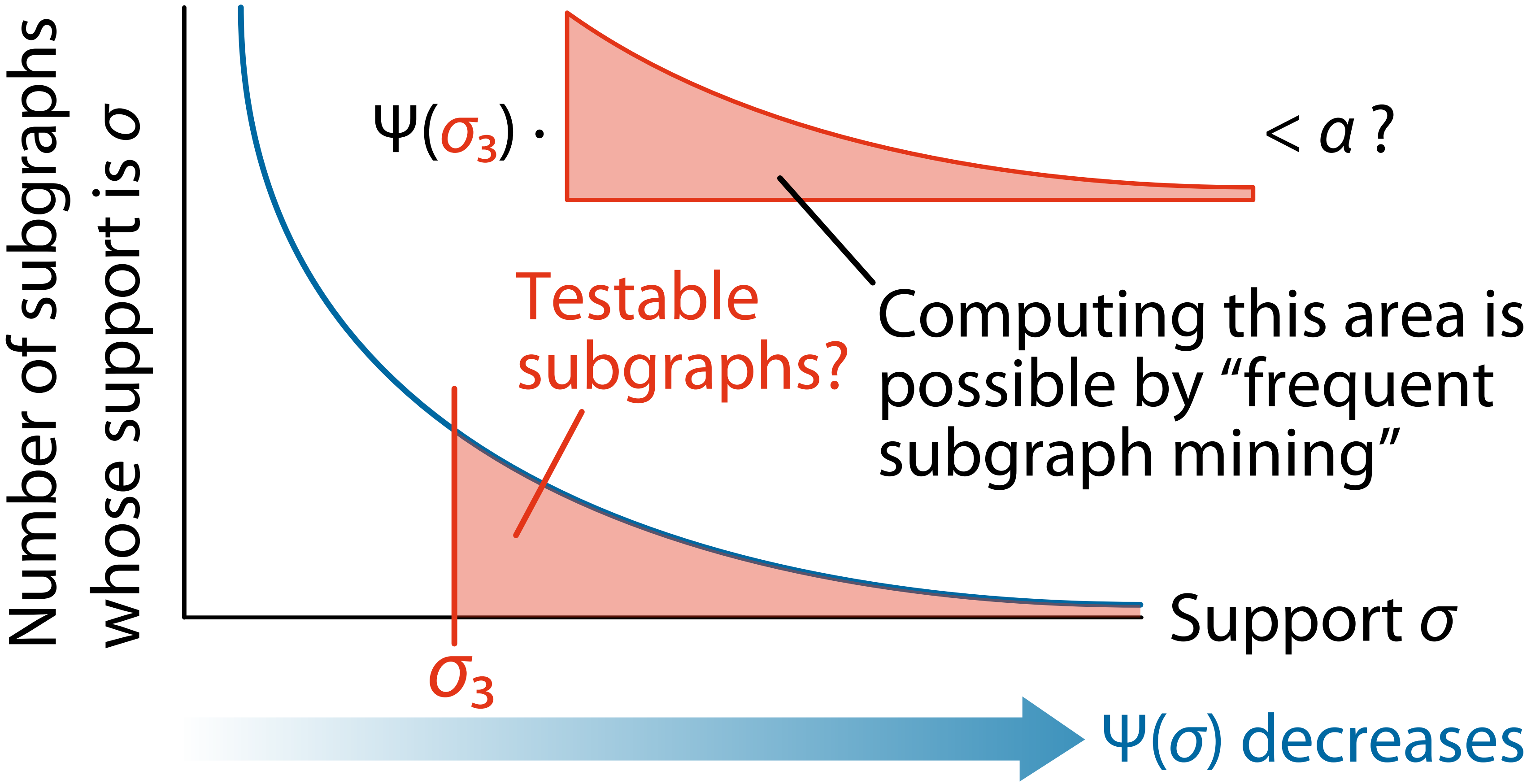
Finding Testable Subgraphs



Finding Testable Subgraphs



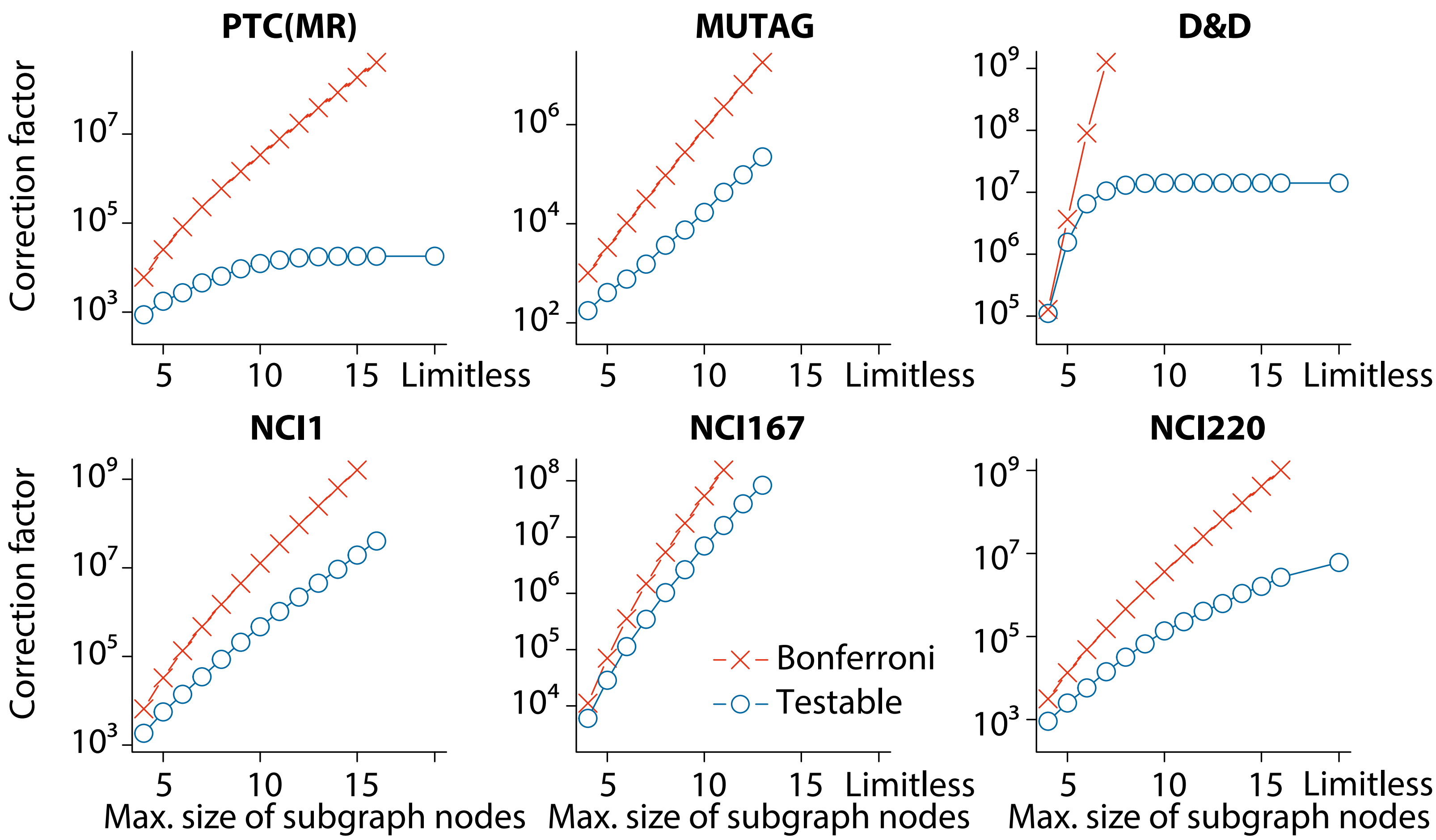
How to Find Testable Subgraphs?



Datasets

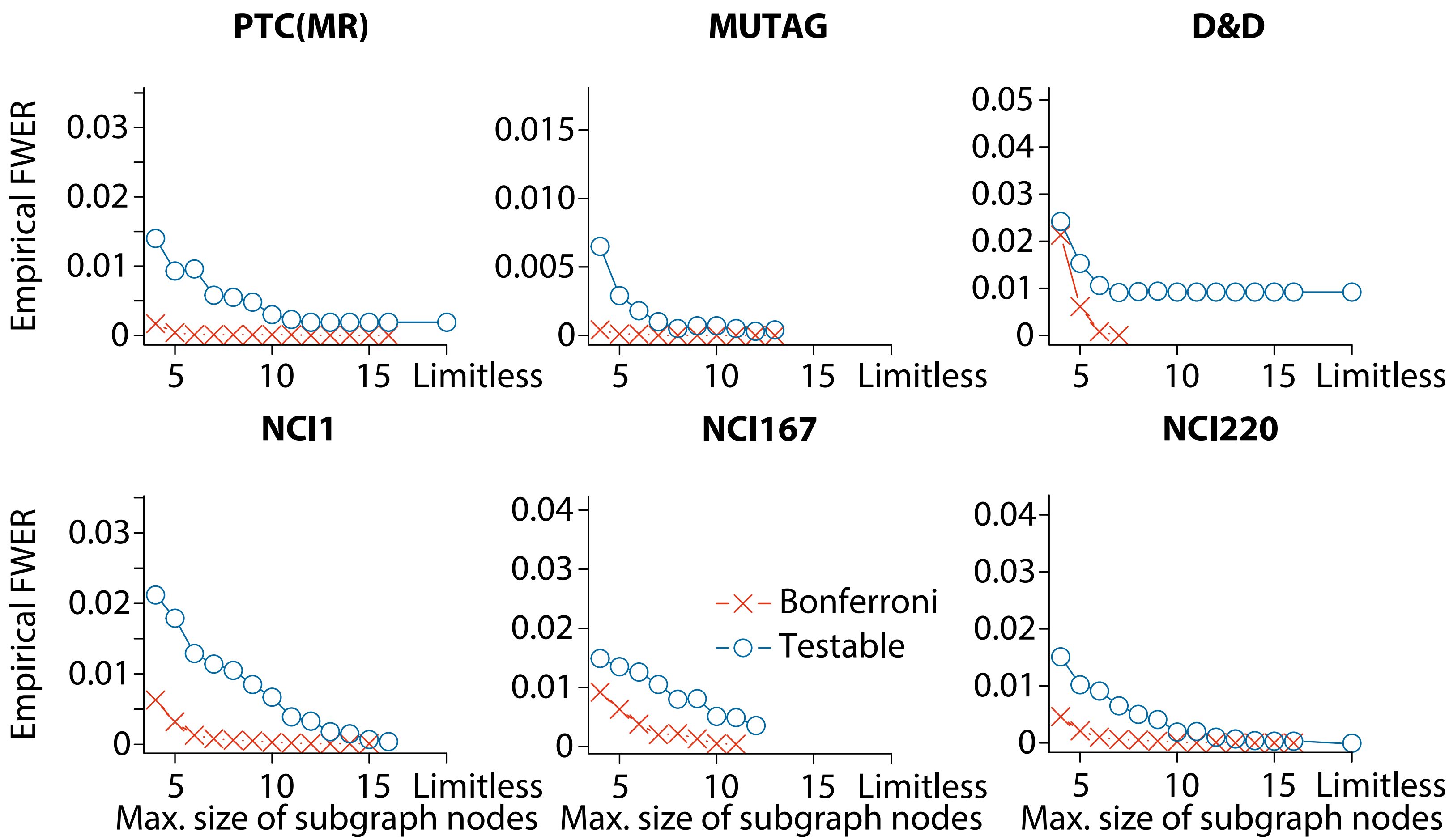
Dataset	Size	#positive	avg.$|V|$	avg.$|E|$	max$|V|$	max$|E|$
PTC (MR)	584	181	31.96	32.71	181	181
MUTAG	188	125	17.93	39.59	28	66
D&D	1178	691	284.32	715.66	5748	14267
NCI1	4208	2104	60.12	62.72	462	468
NCI167	80581	9615	39.70	41.05	482	478
NCI220	900	290	46.87	48.52	239	255

Testable Subgraphs



from [Sugiyama et al. SDM2015]

FWER Is Still Too Low!



from [Sugiyama et al. SDM2015]

Take Dependencies into Account

- **Problem:** Dependencies between subgraphs are not considered
- **Solution:** Permutation test
  - Repeat random permutation of class labels ($10^3 \sim 10^4$ times)
  - Get the null distribution of *p*-values
  - The optimal correction factor can be obtained

Westfall-Young Permutation

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

Westfall-Young Permutation

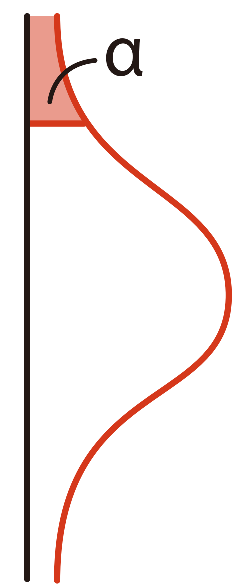
Subgraphs (Hypotheses)

Permutation	H_1	H_2	H_3	…	H_m
1	p_{11}	p_{12}	p_{13}	…	p_{1m}
2	p_{21}	p_{22}	p_{23}	…	p_{2m}
3	p_{31}	p_{32}	p_{33}	…	p_{3m}
⋮	⋮	⋮	⋮	⋱	⋮
h	p_{h1}	p_{h2}	p_{h3}	…	p_{hm}

⇒

Sort and find α-quantile

p^1_{min}
p^2_{min}
p^3_{min}
⋮
p^h_{min}



Using Support for Estimating FWER

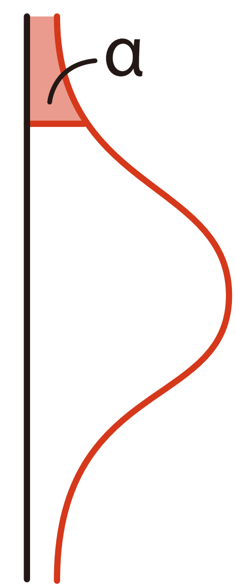
Subgraphs (Hypotheses)

Support

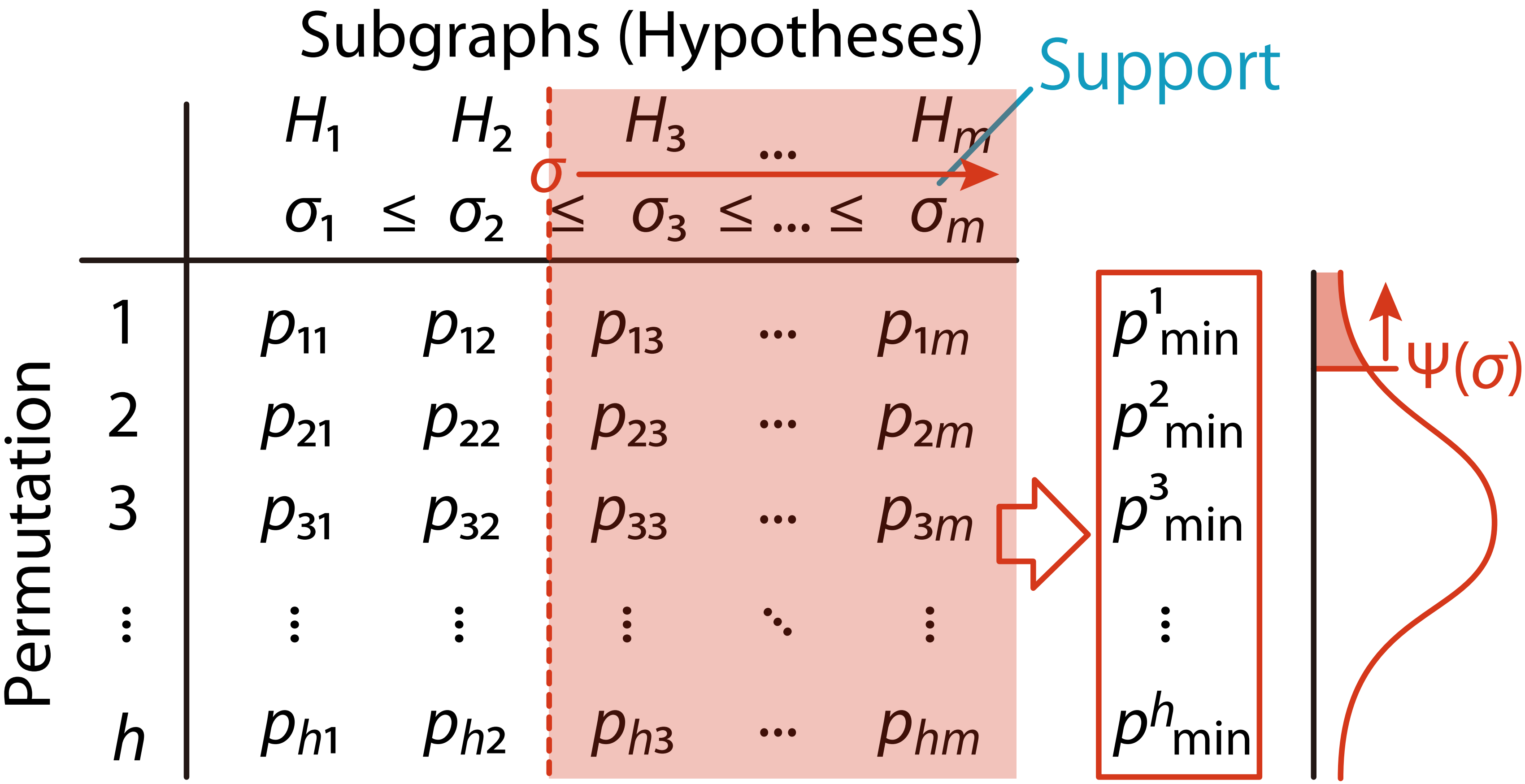
Permutation	H_1	H_2	H_3	...	H_m
	$\sigma_1 \leq$	$\sigma_2 \leq$	$\sigma_3 \leq$	$\ldots \leq$	σ_m
1	p_{11}	p_{12}	p_{13}	...	p_{1m}
2	p_{21}	p_{22}	p_{23}	...	p_{2m}
3	p_{31}	p_{32}	p_{33}	...	p_{3m}
⋮	⋮	⋮	⋮	⋱	⋮
h	p_{h1}	p_{h2}	p_{h3}	...	p_{hm}

Sort and find α-quantile

$p^1_{\min}$
$p^2_{\min}$
$p^3_{\min}$
⋮
$p^h_{\min}$



Estimating FWER

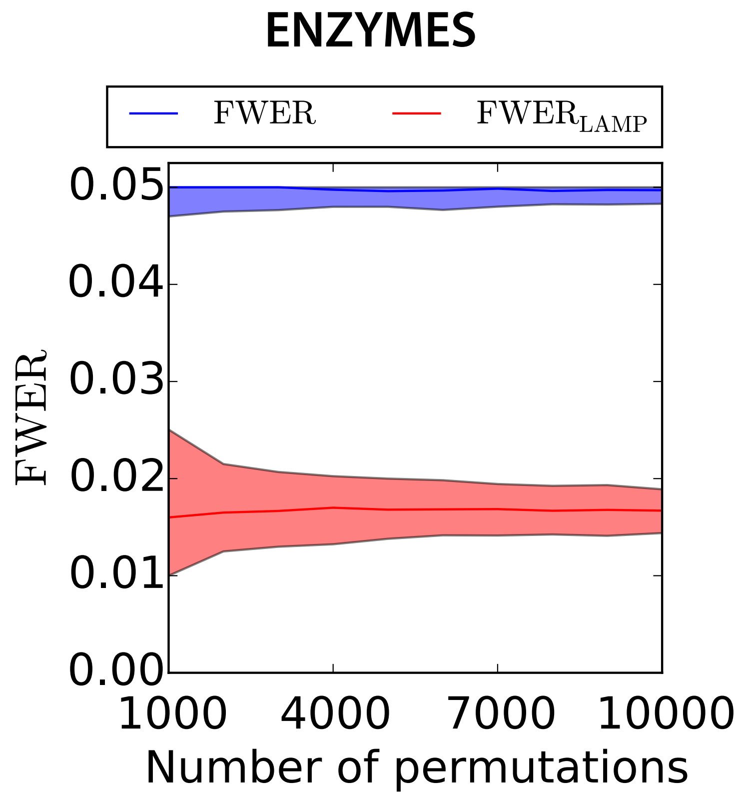


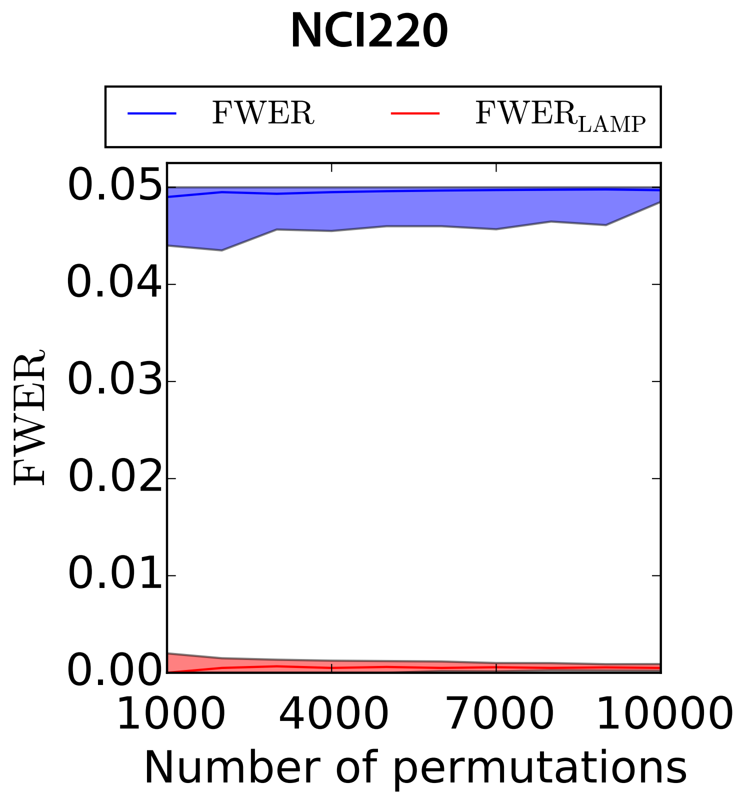
Estimator of FWER $= |\{i : p^i_{\min} \leq \Psi(\sigma)\}| \, / \, h$

"Westfall-Young light" [Llinares-López et al. KDD'15]

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

FWER in Subgraph Mining





from [Llinares-López et al. KDD2015]

Conclusion

- Significant subgraph mining is introduced
  - Find statistically significant subgraphs while controlling the FWER
  - pattern mining (data mining) + MCP (statistics)
    - Sugiyama, M., Llinares-López, F., Kasenburg, N., Borgwardt, K.: **Significant Subgraph Mining with Multiple Testing Correction**, SIAM SDM 2015
    - Llinares-López, F., Sugiyama, M., Papaxanthos, L., Borgwardt, K.: **Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing**, ACM SIGKDD 2015
- Ongoing projects:
  - Find significant subgraphs on a single massive graph
  - Find significant subtrees on a tree