February 16, 2015

Tokyo Workshop on Statistically Sound Data Mining

Multiple Testing Correction in Graph Mining

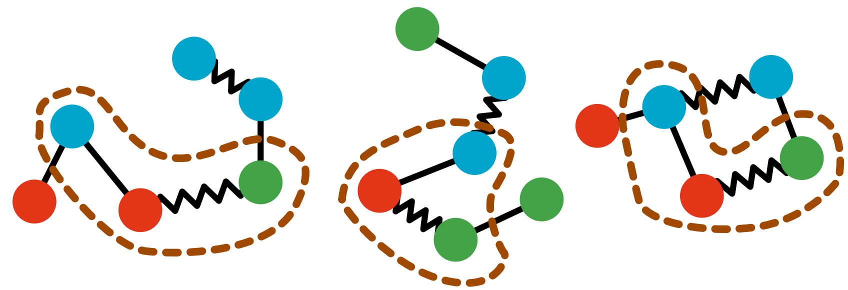
Mahito Sugiyama (Osaka University, JST PRESTO)

Joint work with Felipe Llinares López[1], Niklas Kasenburg[2], Karsten Borgwardt[1] ([1]ETH Zürich, [2]Univ. Copenhagen)

Binary data

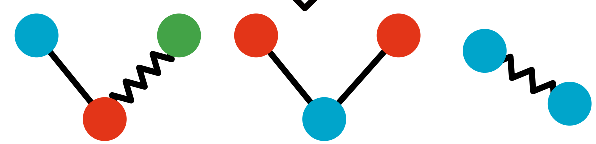
ID	a	b	c	d	e	f	g	h	i	j
1	0	0	1	1	0	0	1	1	1	0
2	1	1	0	1	1	0	1	1	1	0
3	1	0	1	1	0	0	1	1	1	0
4	1	1	0	0	1	0	0	1	0	1
5	1	1	0	1	1	0	1	1	1	0

Graphs



Pattern mining

{a, b, e} {d, g, h, i}

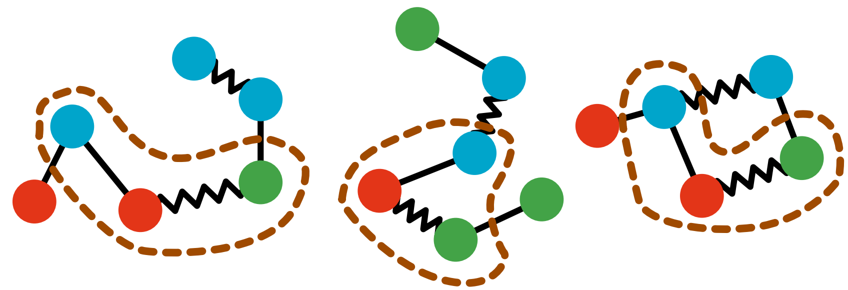


Support: 3 4 3 2 3

Binary data

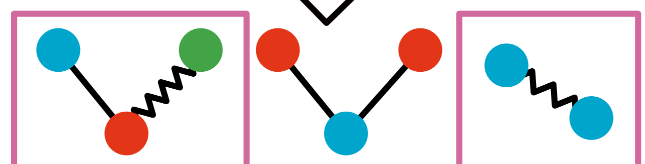
ID	a	b	c	d	e	f	g	h	i	j
1	0	0	1	1	0	0	1	1	1	0
2	1	1	0	1	1	0	1	1	1	0
3	1	0	1	1	0	0	1	1	1	0
4	1	1	0	0	1	0	0	1	0	1
5	1	1	0	1	1	0	1	1	1	0

Graphs



Pattern mining

{a, b, e} {d, g, h, i}



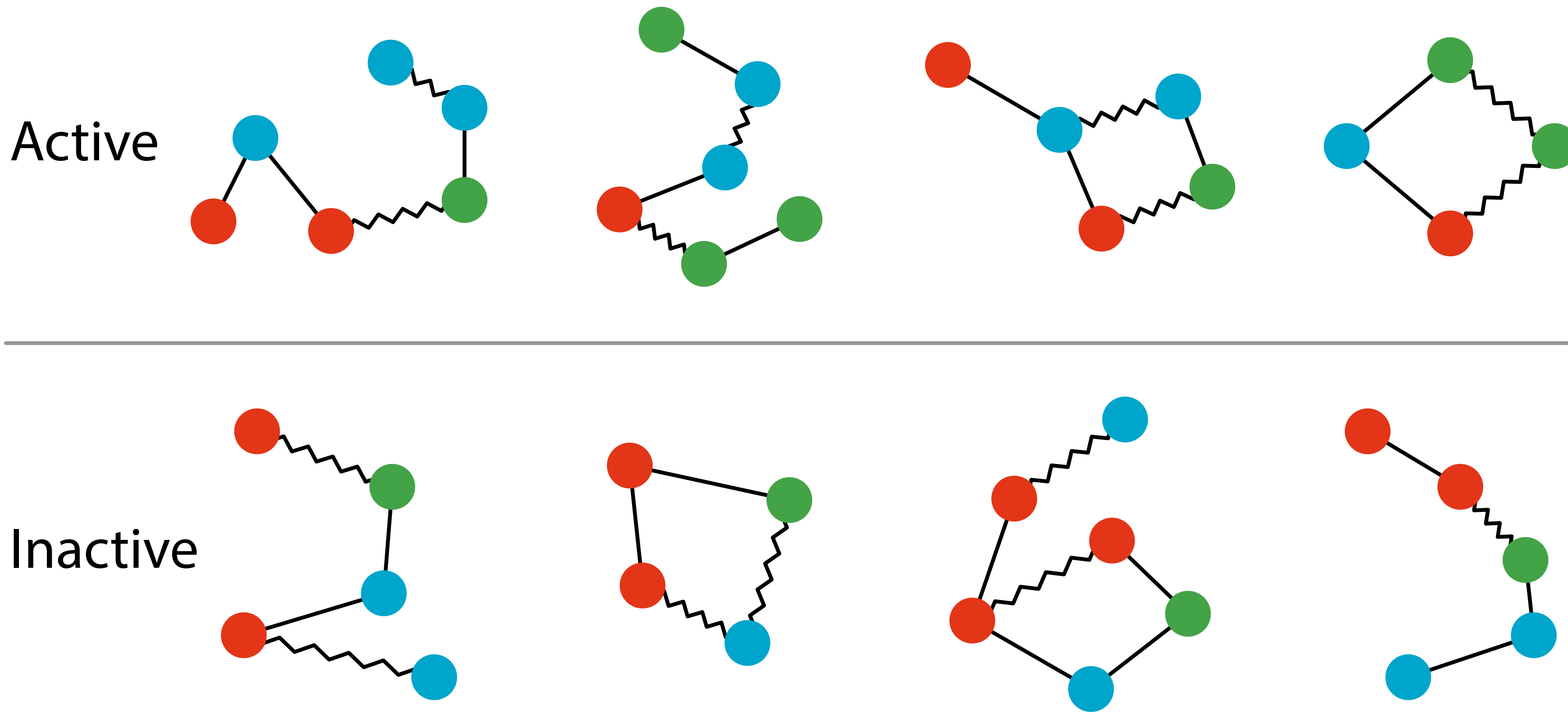
Support: 3 4 3 2 3

(Statistically) Significant patterns
(P value < 0.05)

P value: 0.06 0.01 0.02 0.07 0.02

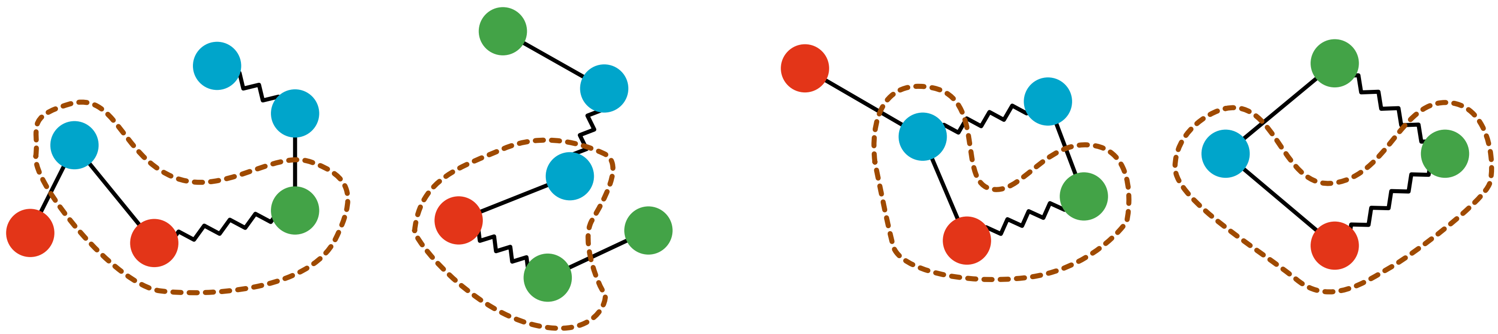
The P value is crucial for scientific discovery!

Find Subgraphs

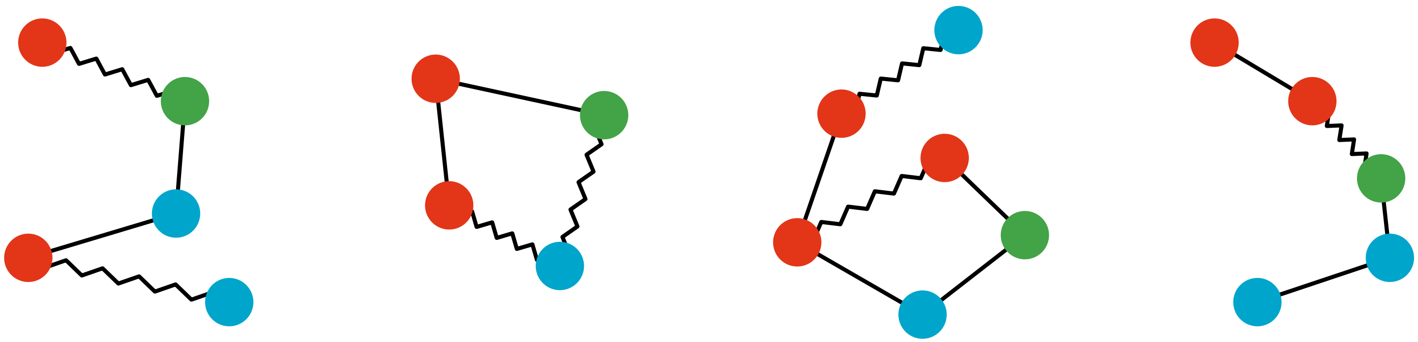


Find Subgraphs

Active

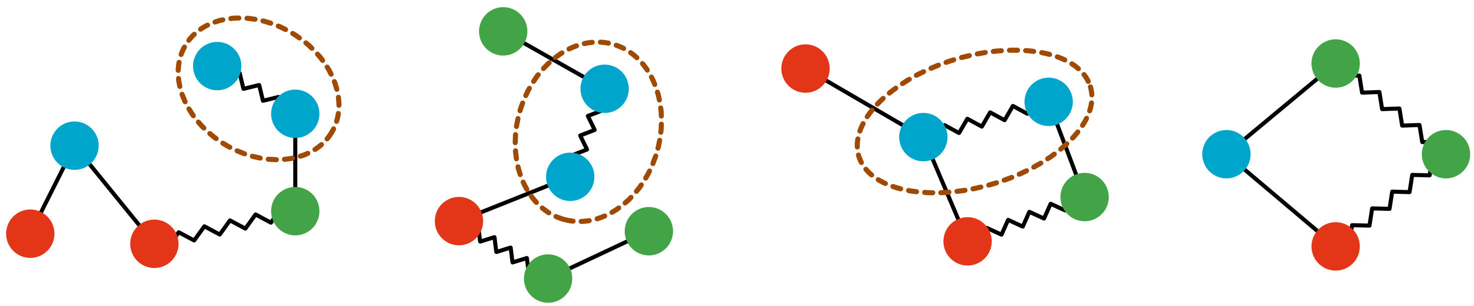


Inactive

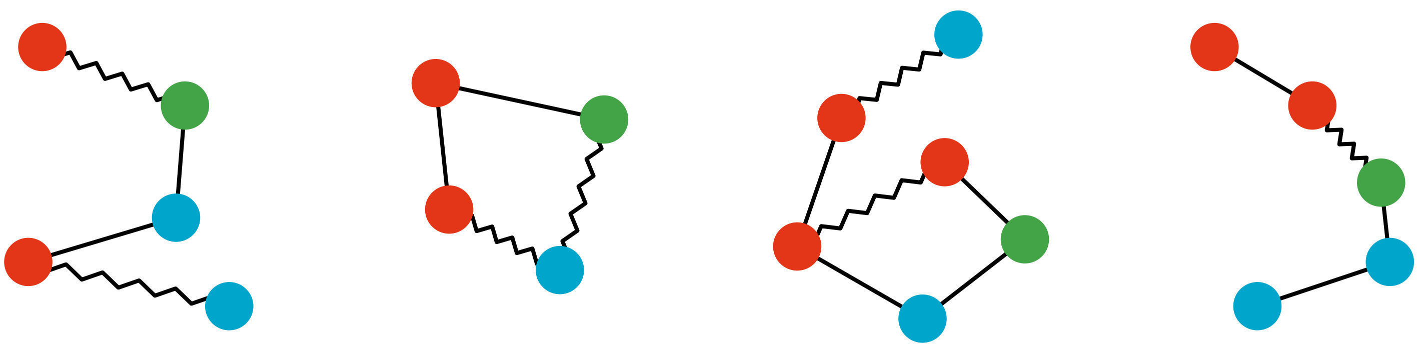


Find Subgraphs

Active

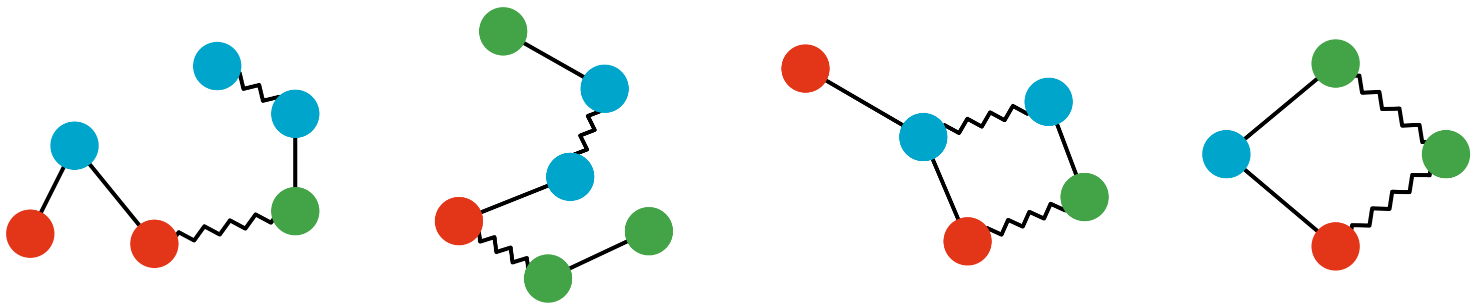


Inactive

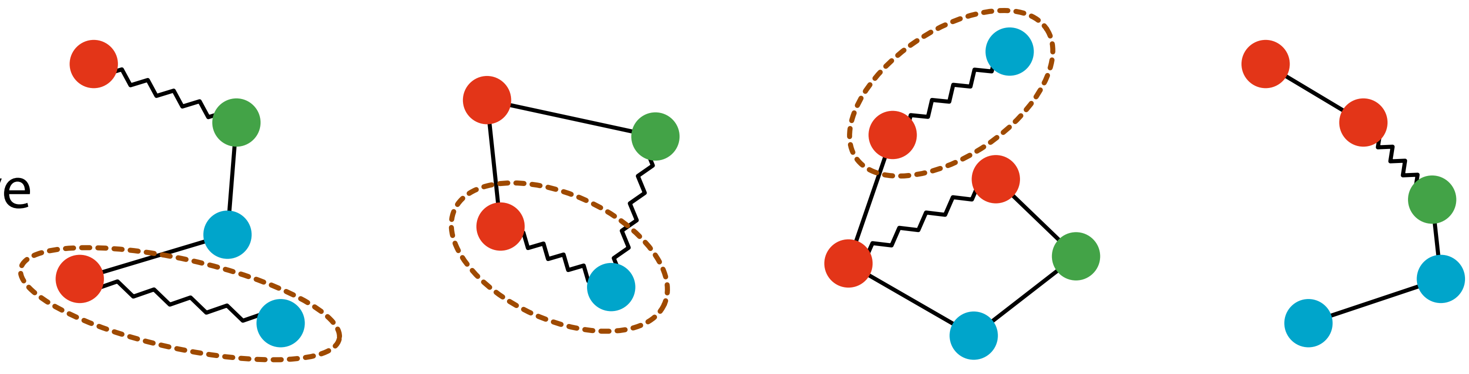


Find Subgraphs

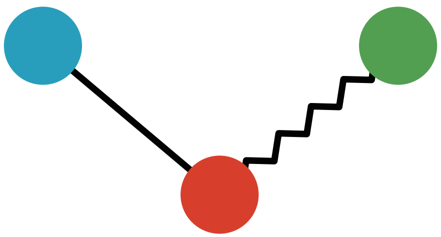
Active



Inactive



Hypothesis Test for Each Subgraph



	Alternative hypothesis is true	Null hypothesis is true
Declared significant	True Positive	False Positive (Type I Error)
Declared non-significant	False Negative (Type II Error)	True Negative

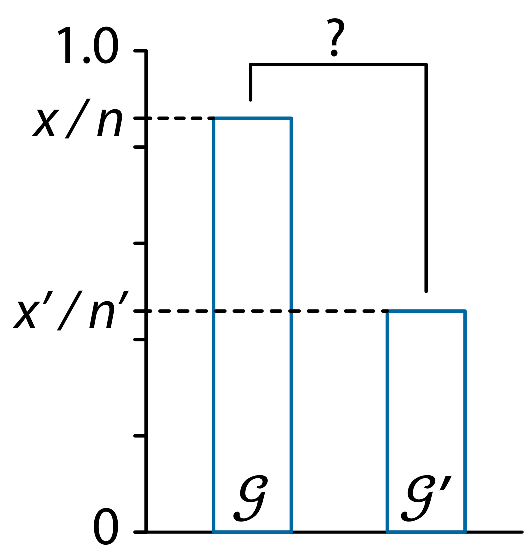
Null hypothesis: The occurence of the subgraph is independent from the activity

Alternative hypothesis: The occurence of the subgraph is associated with the activity

Testing the Independence of Subgraph

- Given two sets of graphs $\mathcal{G}$ and $\mathcal{G}'$
  - $|\mathcal{G}| = n, |\mathcal{G}'| = n'$ $(n \leq n')$
- The *P* value of each subgraph $H \sqsubseteq G$ with $G \in \mathcal{G} \cup \mathcal{G}'$ is determined by the Fisher's exact test

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

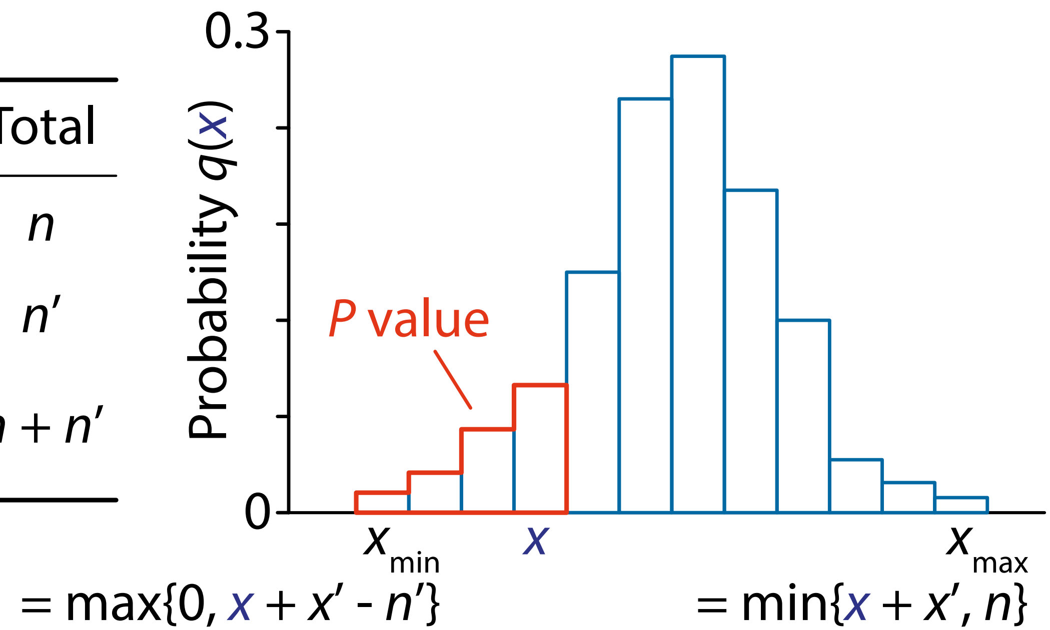


Fisher's Exact Test

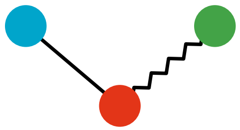
- The probability $q(x)$ of obtaining x and x' is given by the hypergeometric distribution:

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

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



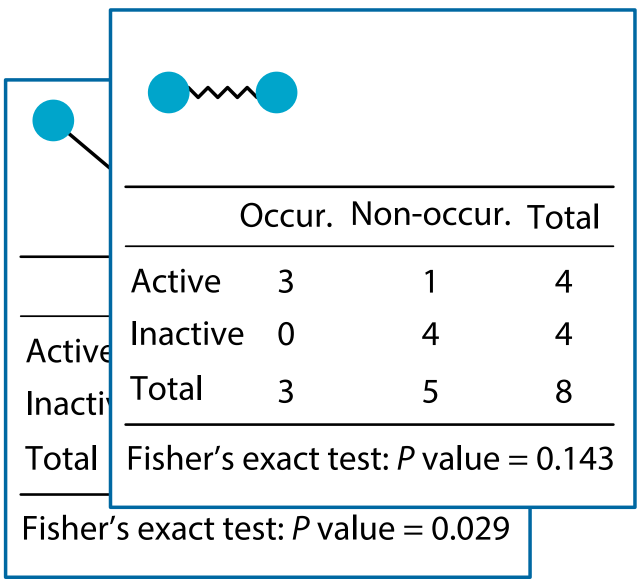
Multiple Testing



	Occur.	Non-occur.	Total
Active	4	0	4
Inactive	0	4	4
Total	4	4	8

Fisher's exact test: P value = 0.029

Multiple Testing



Multiple Testing

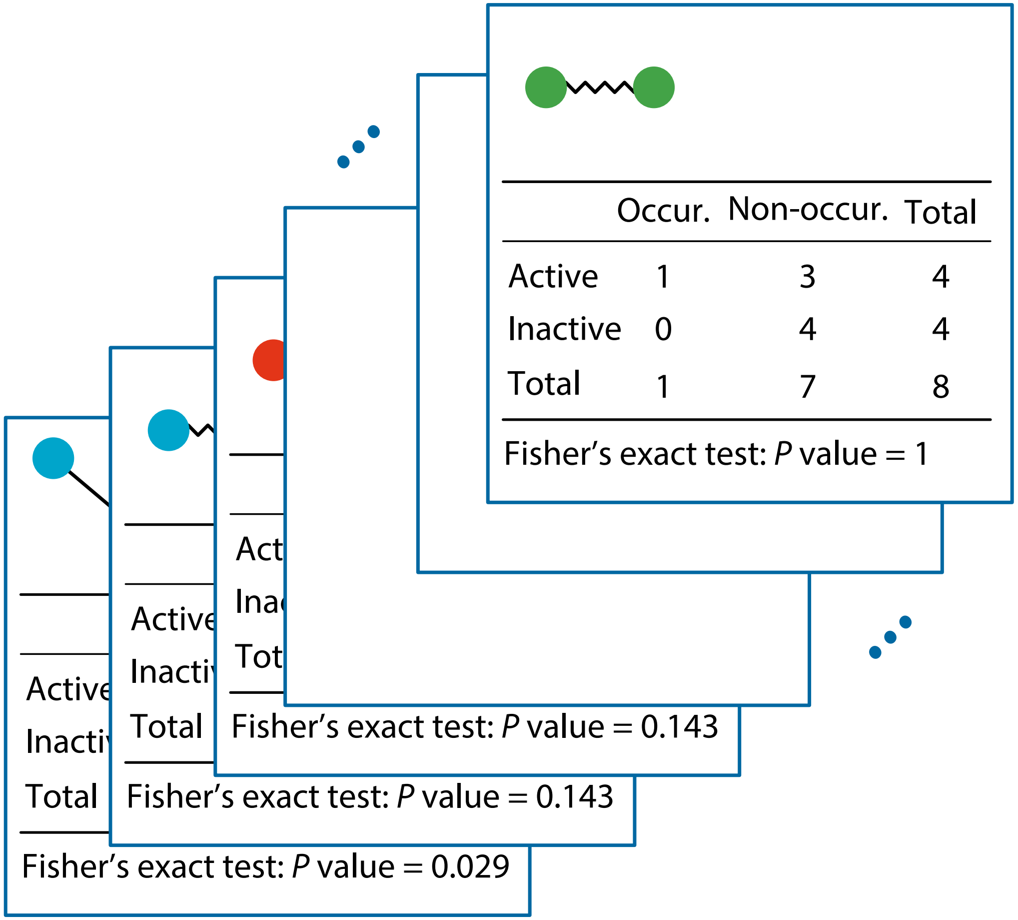
	Occur.	Non-occur.	Total
Active	0	4	4
Inactive	3	1	4
Total	3	5	8

Fisher's exact test: P value = 0.143

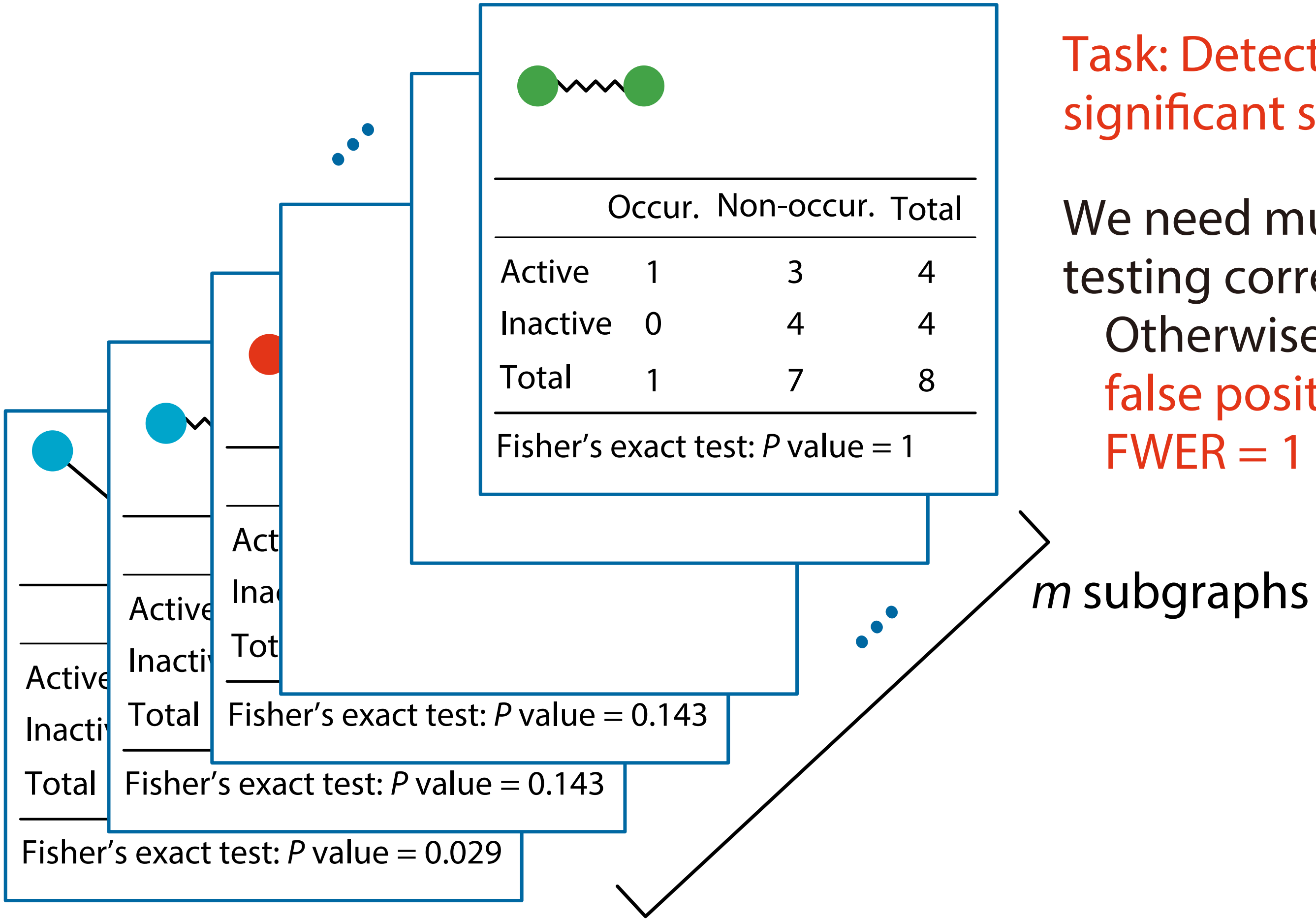
Fisher's exact test: P value = 0.143

Fisher's exact test: P value = 0.029

Multiple Testing



Multiple Testing

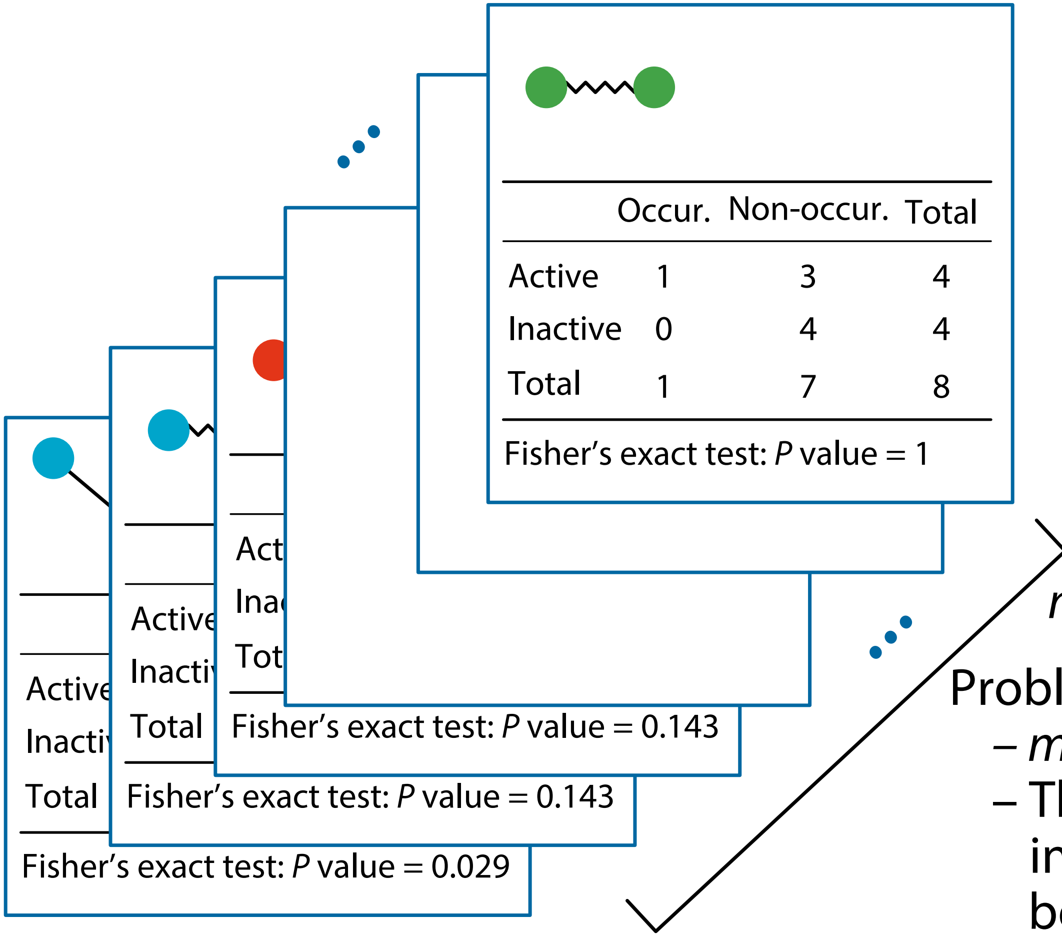


Task: Detect all significant subgraphs

We need multiple testing correction!

Otherwise, too many false positives: $\mathrm{FWER} = 1 - (1 - \alpha)^m$

Multiple Testing



Task: Detect all significant subgraphs

We need multiple testing correction!

Otherwise, too many false positives: $FWER = 1 - (1 - \alpha)^m$

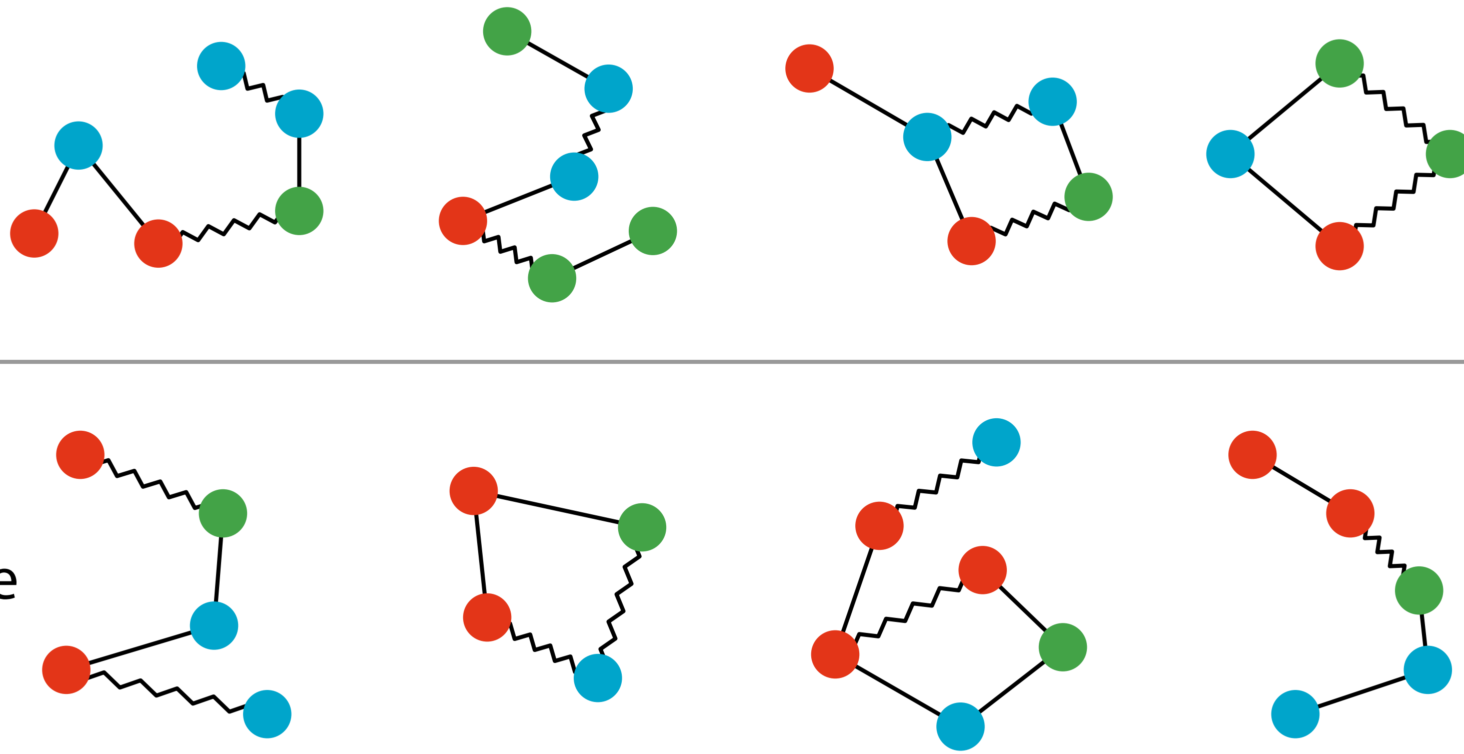
Problems:

- m is massive
- The significance level α / m in Bonferroni correction becomes too conservative

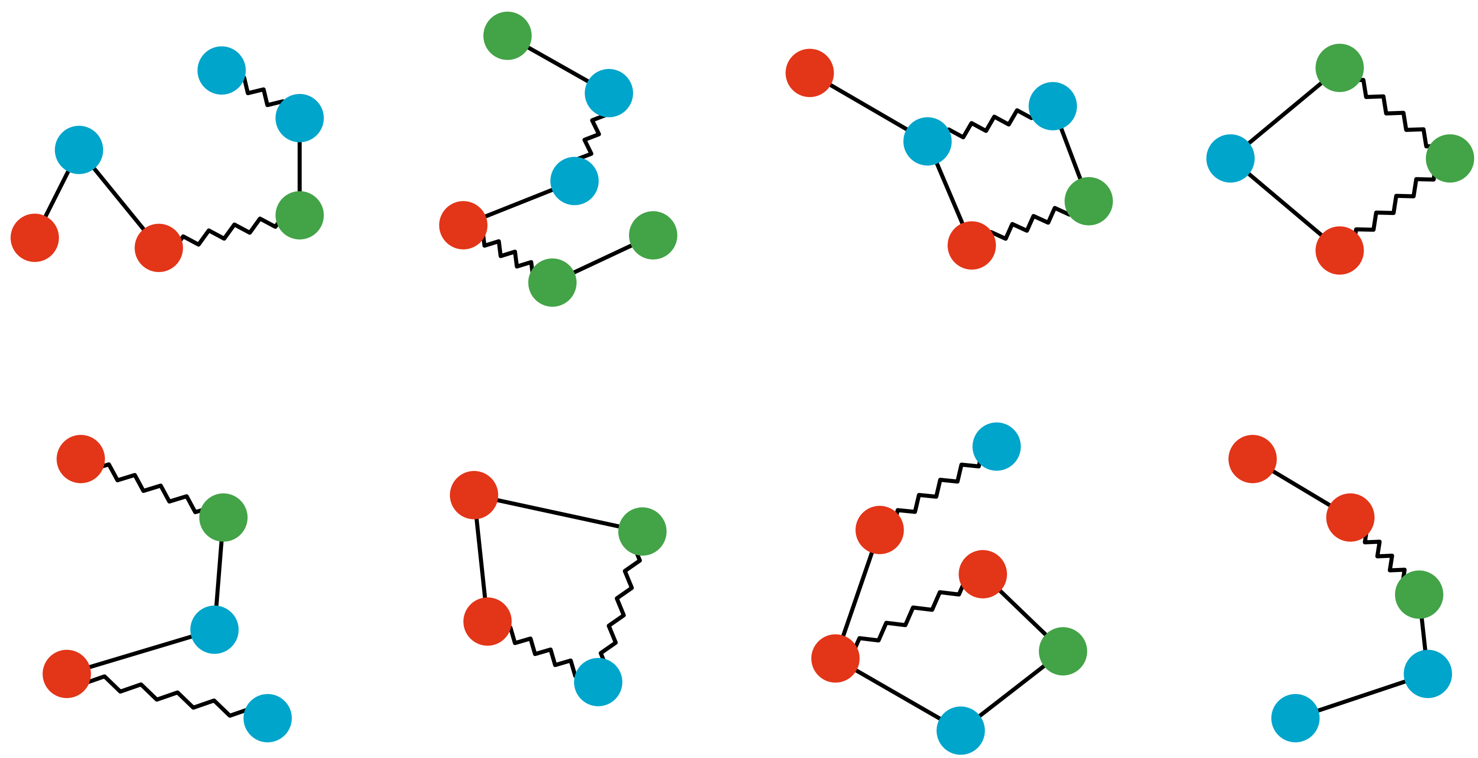
Counting the Frequency of Subgraphs

Active

Inactive

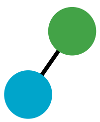


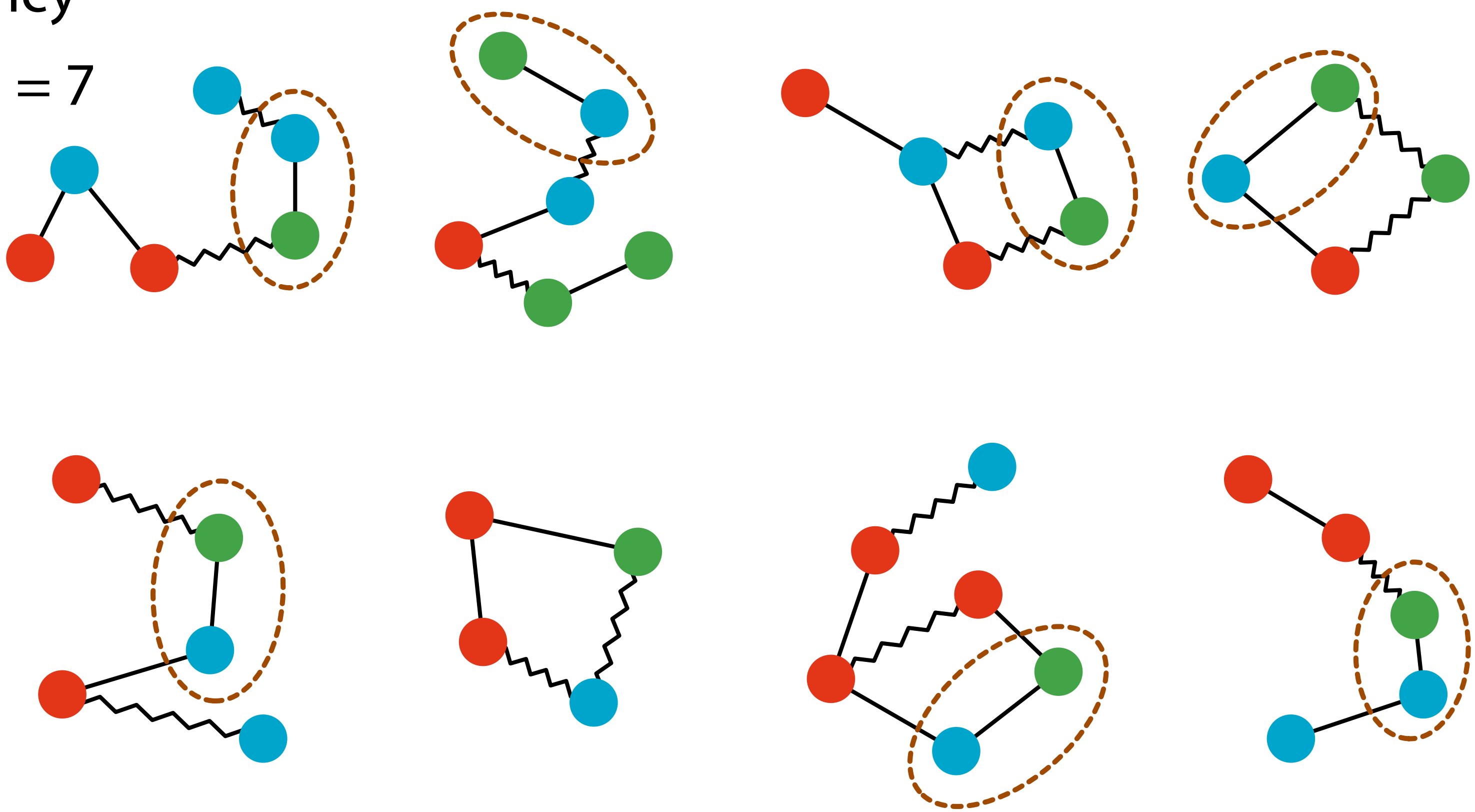
Counting the Frequency of Subgraphs



Counting the Frequency of Subgraphs

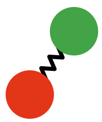
Frequency

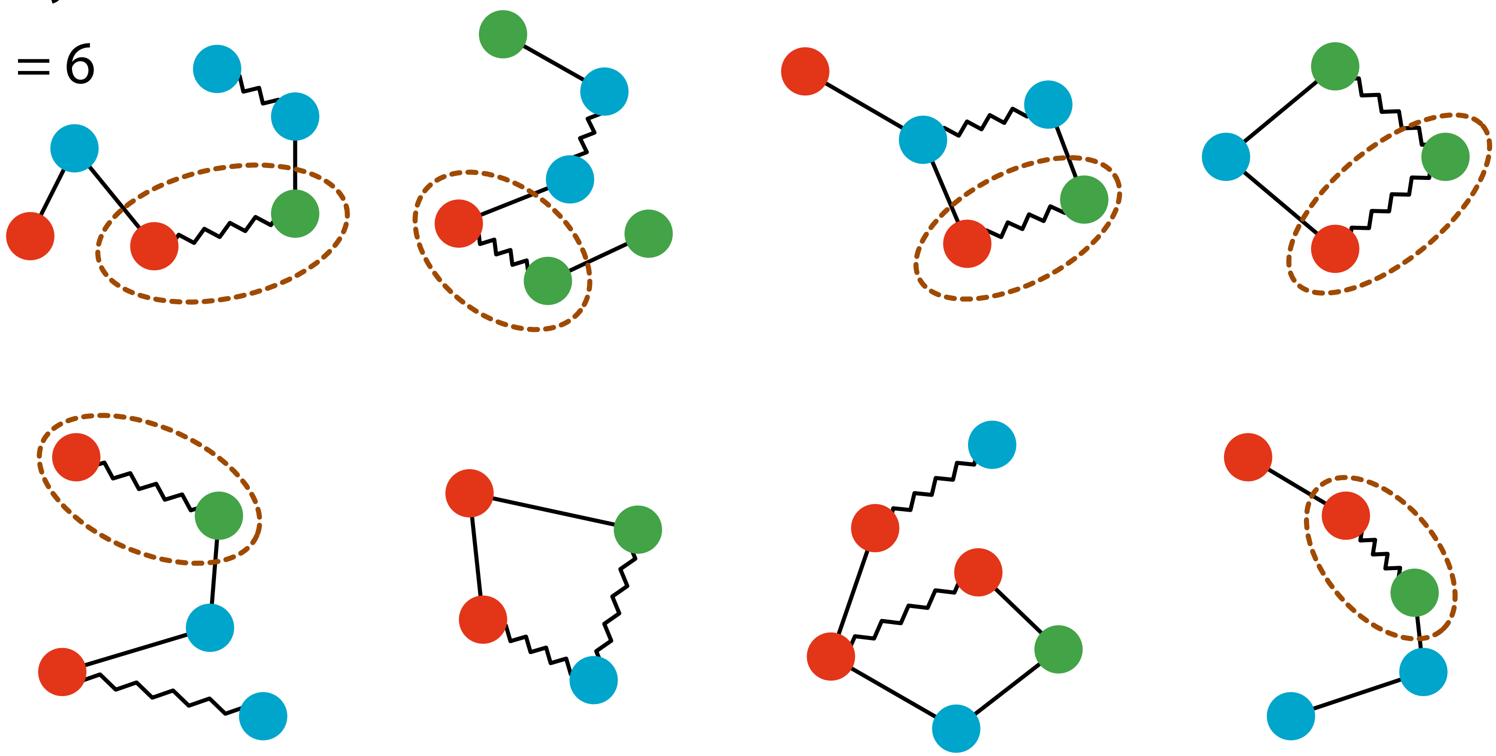
$f($  $) = 7$



Counting the Frequency of Subgraphs

Frequency

$f($  $) = 6$



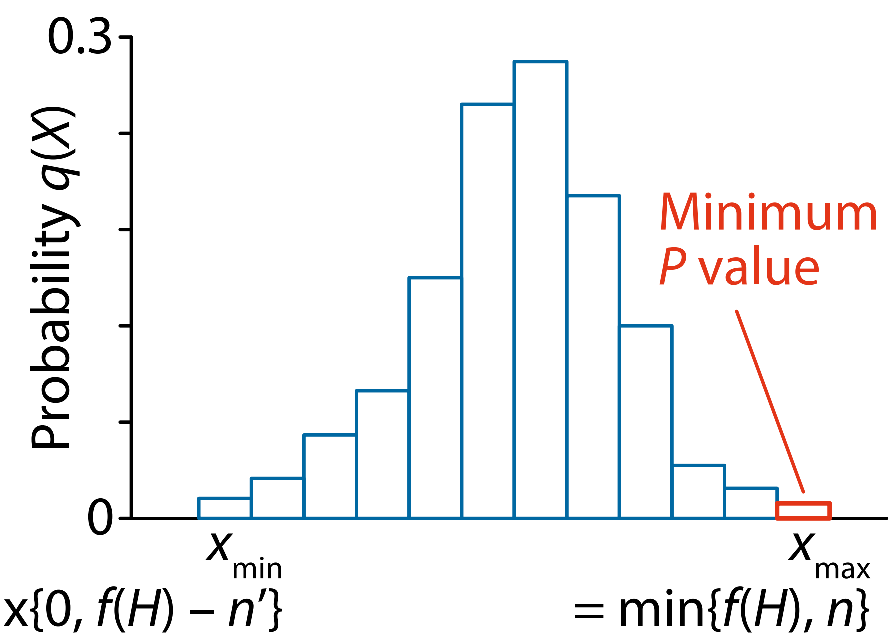
The Minimum *P* Value

- The minimum achievable *P* value for the frequency $f(H)$ of a subgraph H is

$$P_{\min} = \binom{n}{f(H)} \Big/ \binom{n+n'}{f(H)}$$

	Occ.	Non-occ.	Total
Active	$f(H)$	$n - f(H)$	n
Inactive	0	n'	n'
Total	$f(H)$	$(n - f(H)) + n'$	$n + n'$

Most biased case ($f(H) < n$)



Testability

- The minimum achievable P value
  for the frequency $f(H)$ of a subgraph H is

$$P_{\min} = \binom{n}{f(H)} \bigg/ \binom{n + n'}{f(H)}$$

- Tarone (1990) pointed out (and Terada et al. (2013) revisited):
  For a hypothesis H, if its minimum P value is smaller than the significance threshold, this is untestable and we can ignore it
  - Untestable hypotheses (subgraphs) do not increase the FWER
  - The Bonferroni factor reduces to the number of testable hypotheses

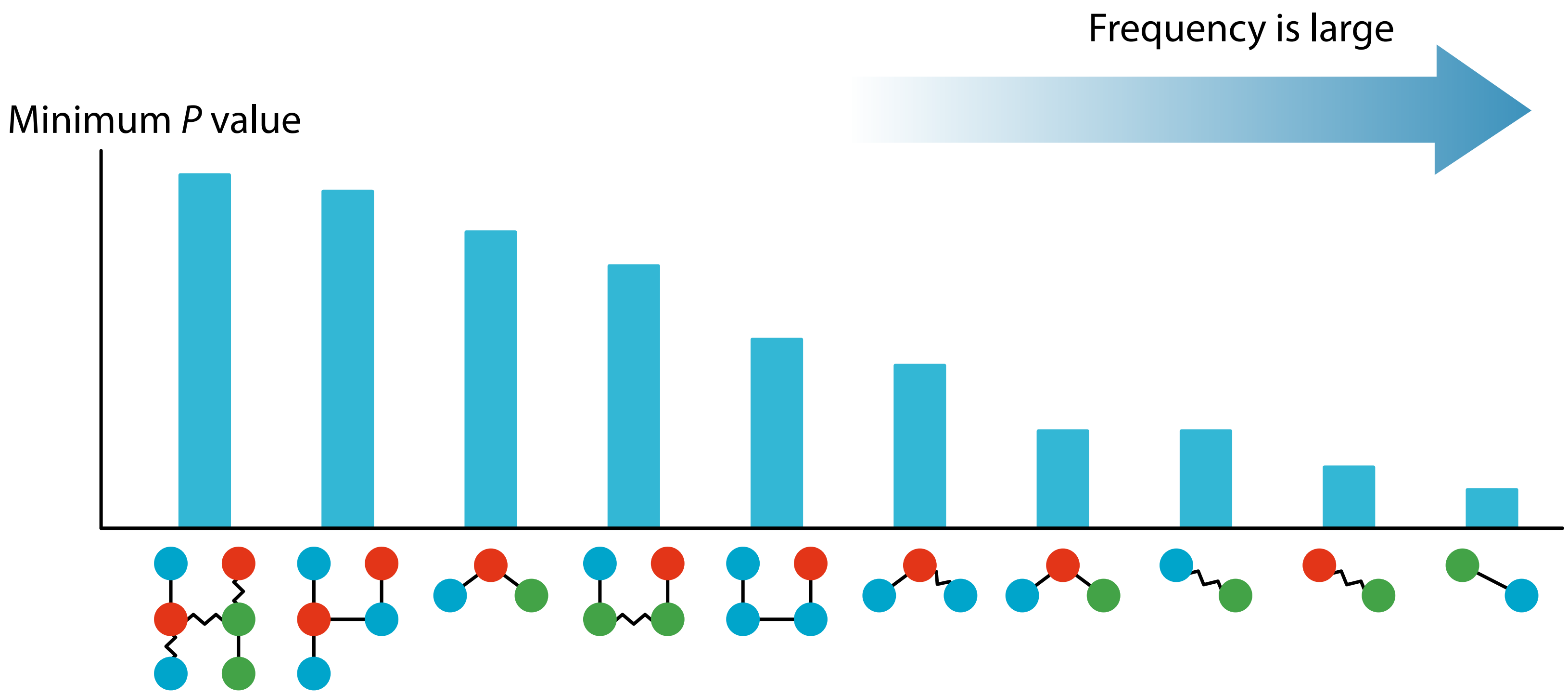
Finding the Optimal Correction Factor

- $m(k)$: # of subgraphs whose minimum P values $< \alpha/k$
  - k: the correction factor, α/k: the corrected significance level
- For each k, FWER is controlled as (Tarone 1990):

$$\mathrm{FWER} \le m(k)\frac{\alpha}{k} = \frac{m(k)}{k}\alpha$$

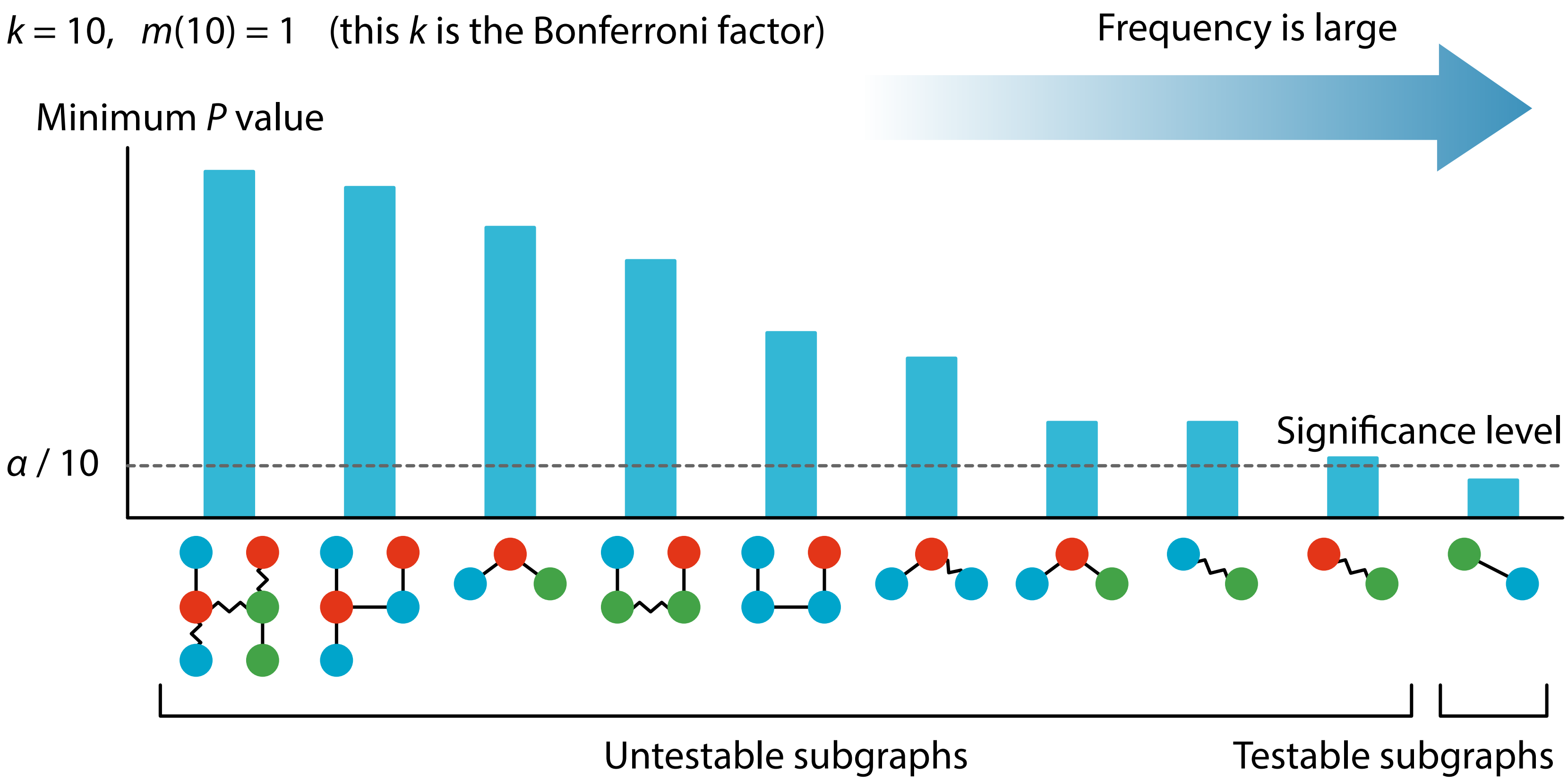
- Our task:
  - Find the smallest k while controlling FWER $\le \alpha$
    - Coincides with the "root" k_{rt} of the function $m(k) - k$
    - $m(k) \le k$ for all $k \ge k_{\mathrm{rt}}$ and $m(k) > k$ for all $k < k_{\mathrm{rt}}$
  - Enumerate testable subgraphs whose min. P values $< \alpha/k_{\mathrm{rt}}$

Testable Subgraphs



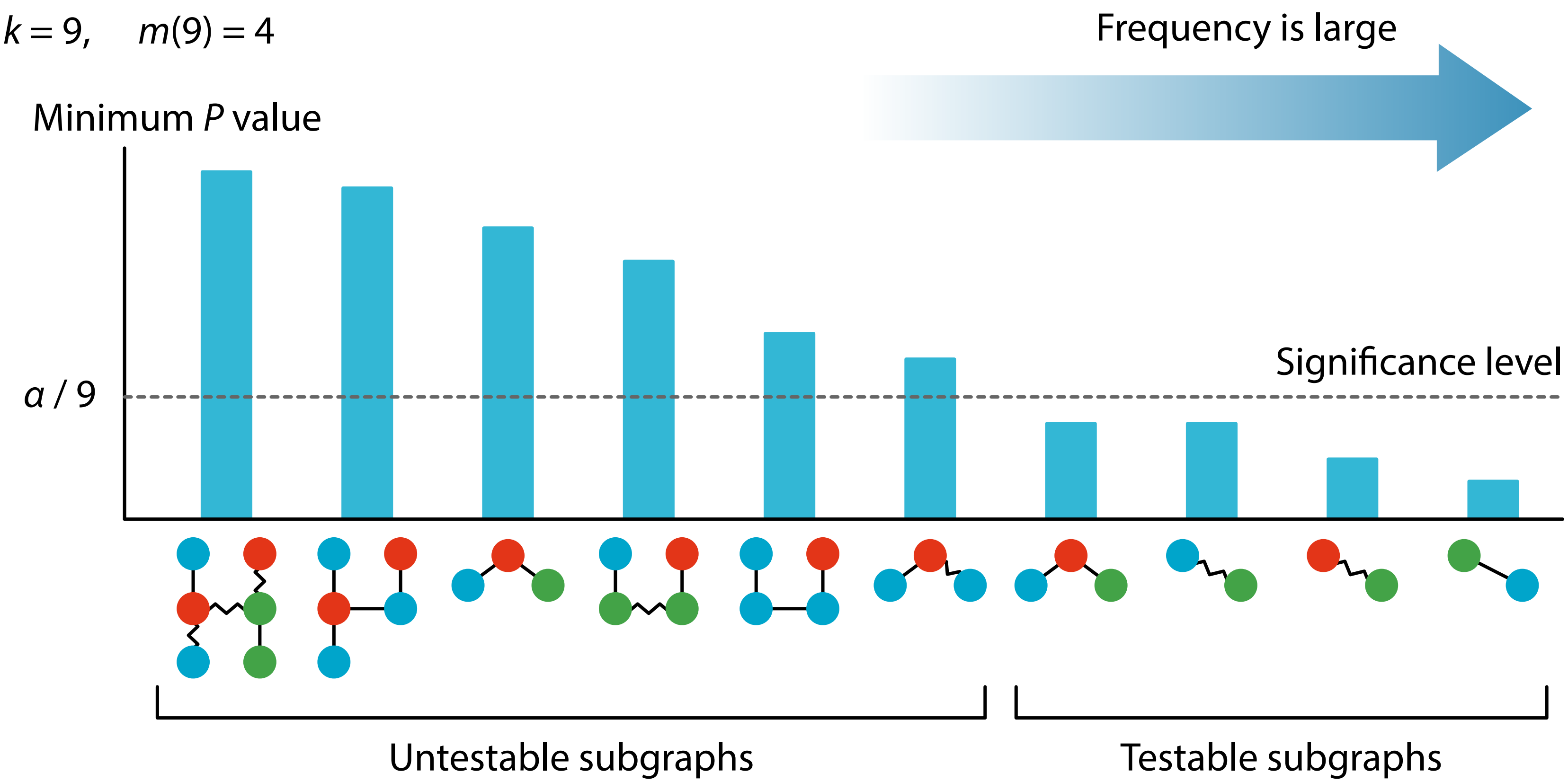
Testable Subgraphs

$k = 10$, $m(10) = 1$ (this k is the Bonferroni factor)

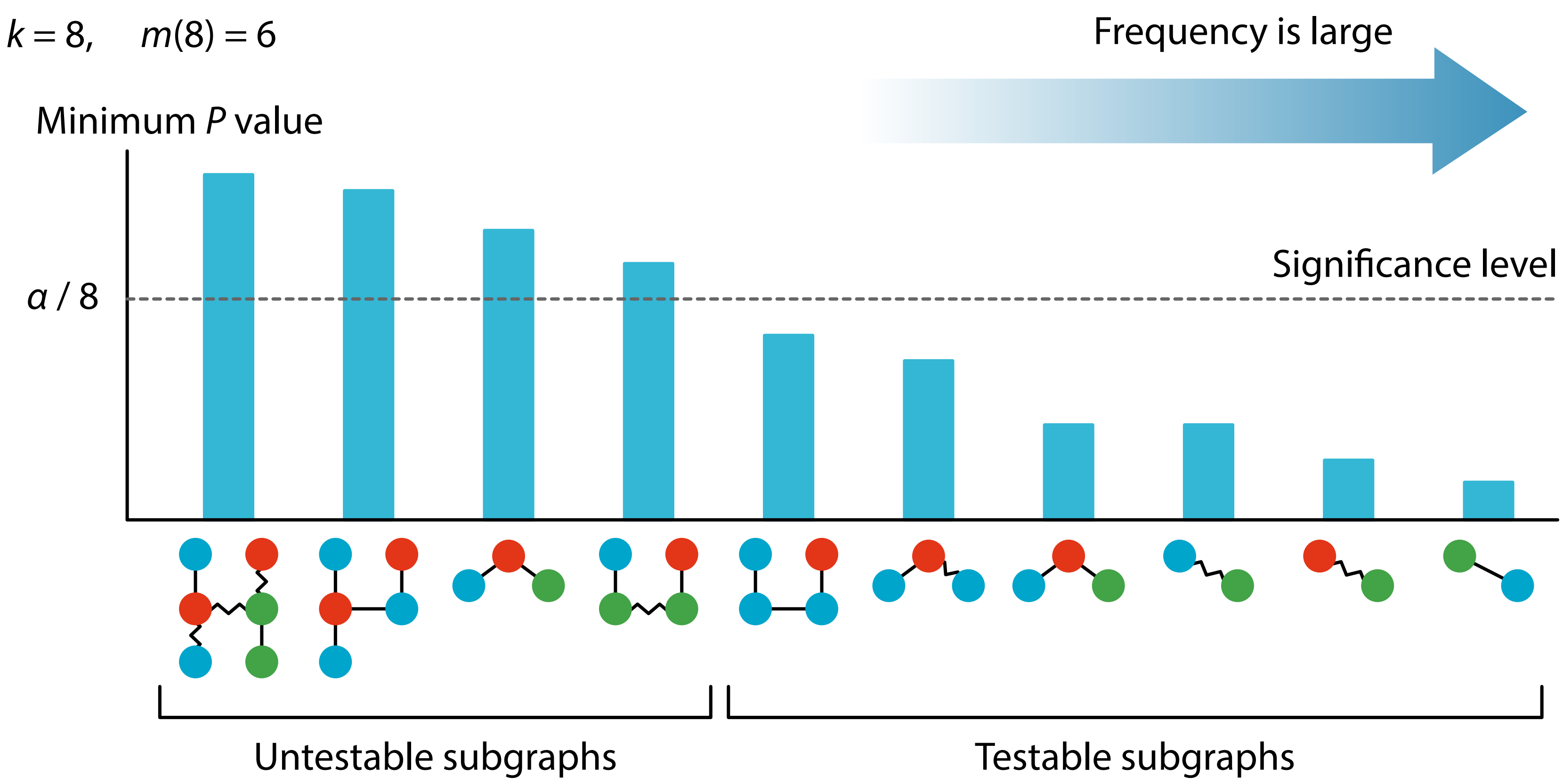


Testable Subgraphs

$k = 9, \quad m(9) = 4$

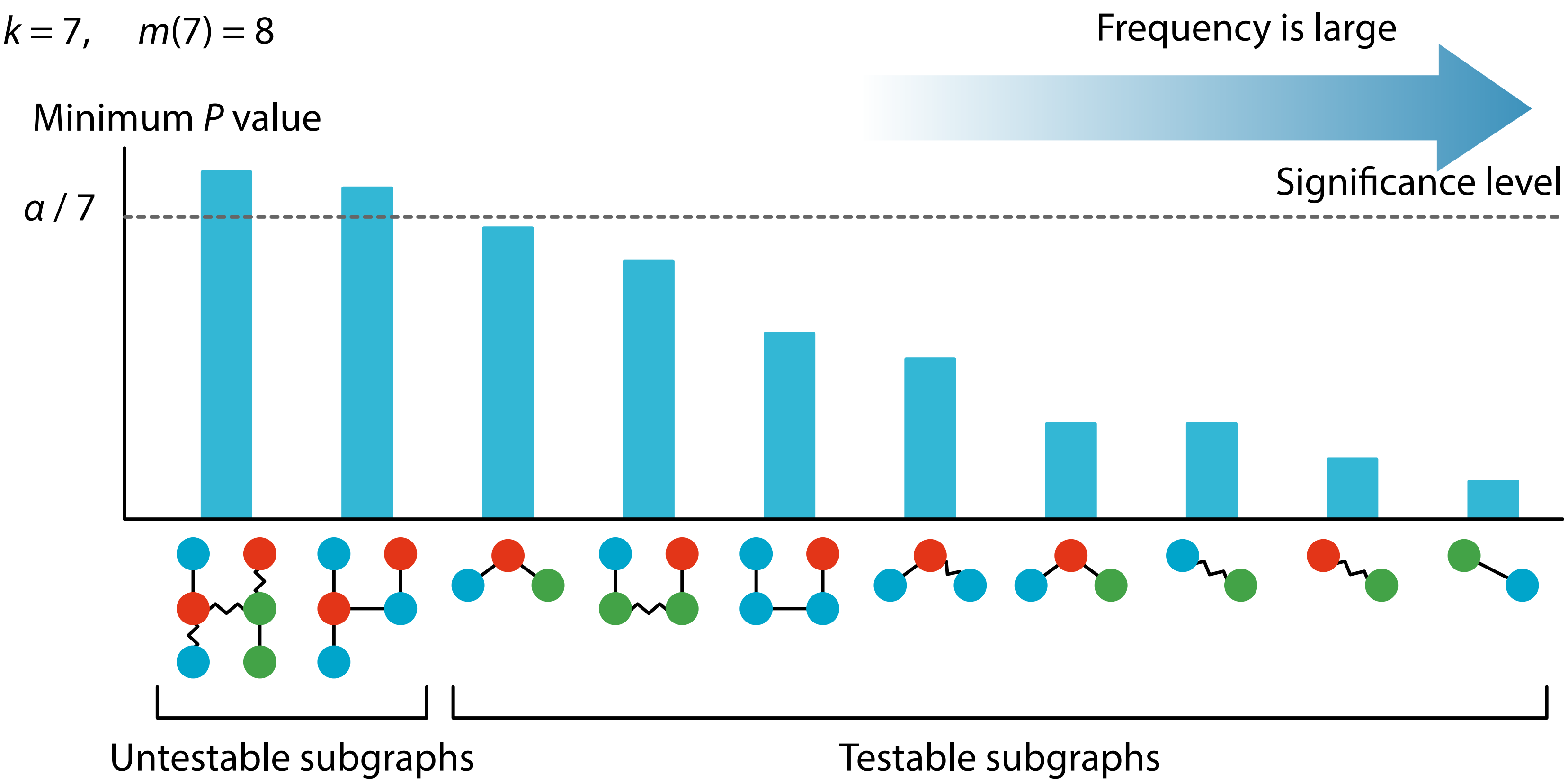


Testable Subgraphs

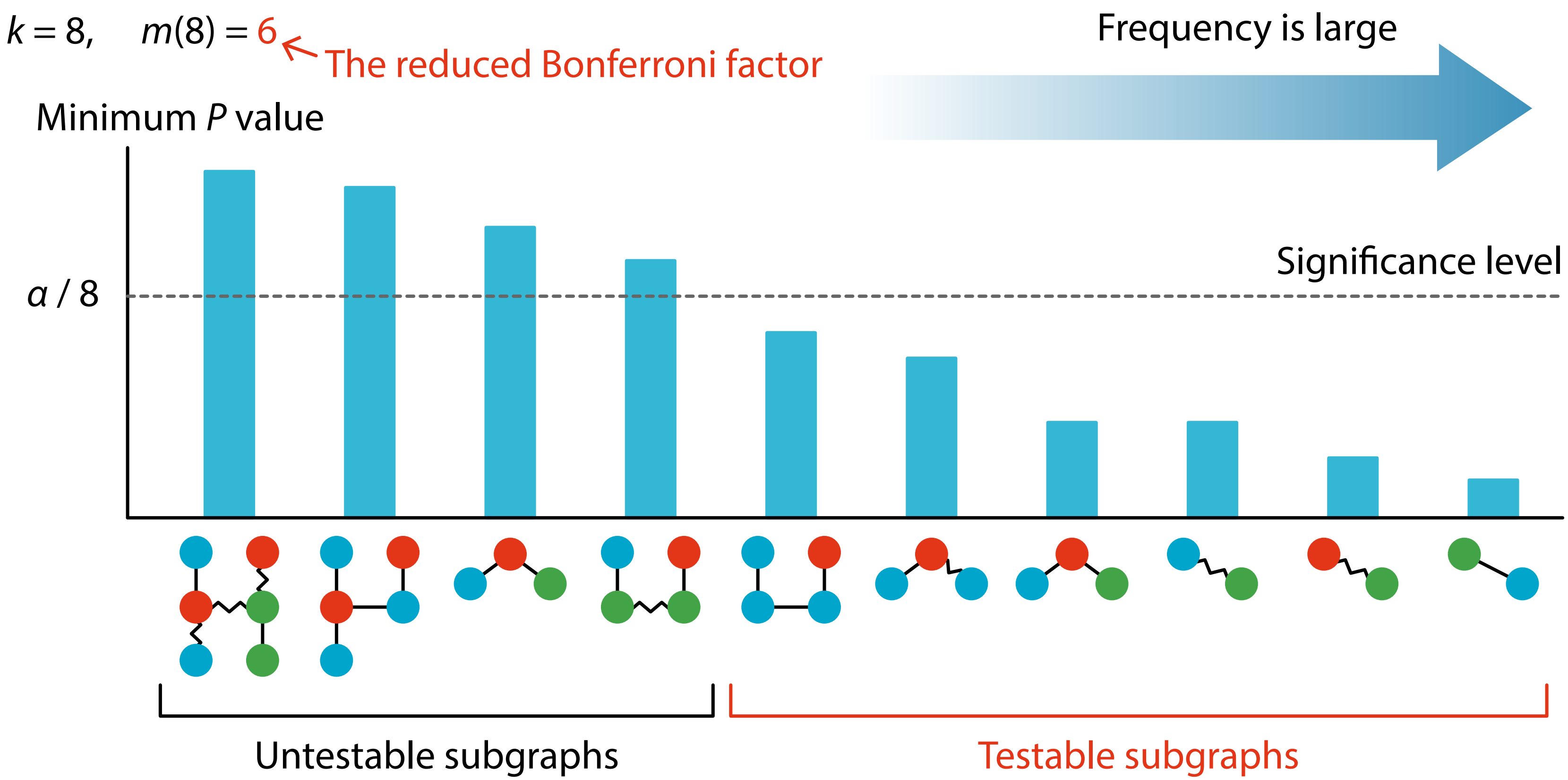


Testable Subgraphs

$k = 7, \quad m(7) = 8$



Testable Subgraphs



Compute the (exact) P values of these testable subgraphs

Subgraphs Are Testable Iff Frequent

- Our task:

Find k such that

(# of subgraphs whose minimum P values $< \alpha / k$) $= k$

Subgraphs Are Testable Iff Frequent

- Our task:

Find k such that

(# of subgraphs whose minimum P values $< \alpha / k$) $= k$

$$\Downarrow$$

Find σ such that

(# of subgraphs whose supports $\geq \sigma$) $= \alpha / \psi(\sigma)$

Subgraphs Are Testable Iff Frequent

- Our task:

Find k such that

(# of subgraphs whose minimum P values $< \alpha/k$) $= k$

$$\Downarrow$$

Find σ such that

(# of subgraphs whose supports $\geq \sigma$) $= \alpha/\psi(\sigma)$

Testable subgraphs = Frequent subgraphs

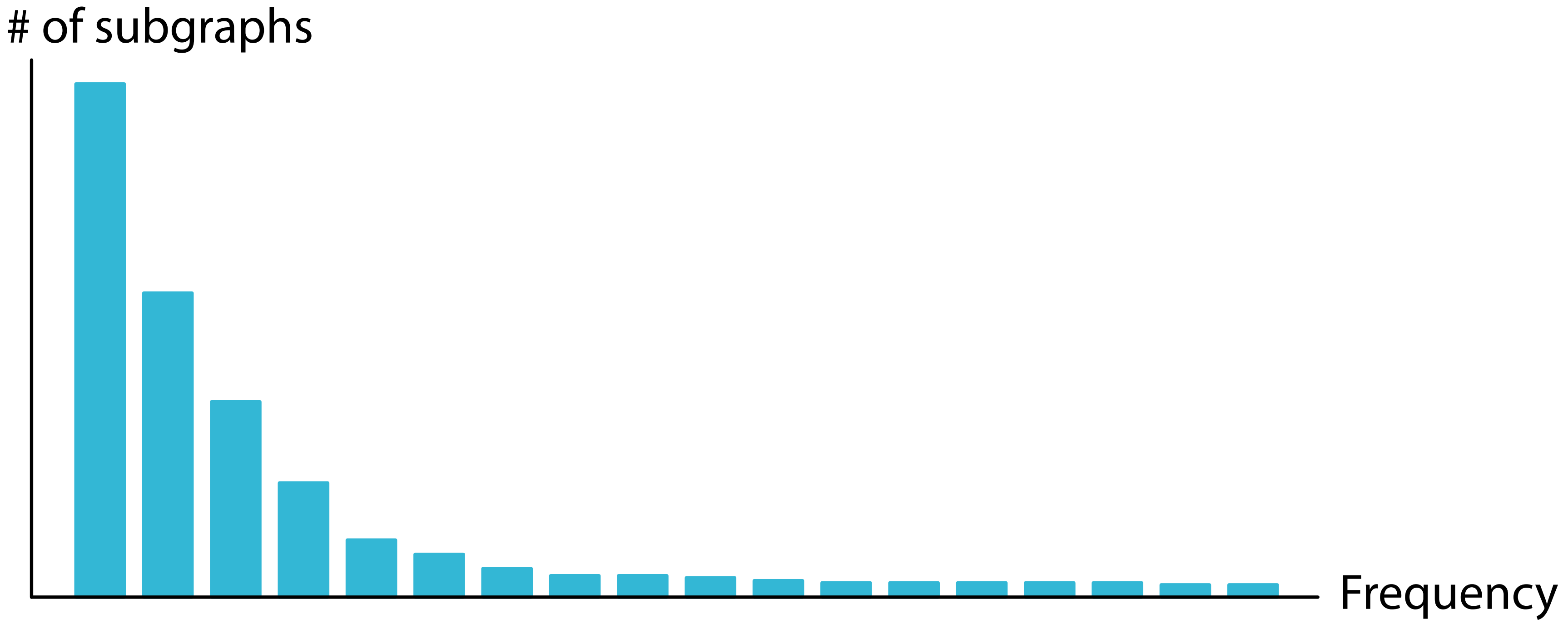
Use Frequent Subgraph Mining

- Testable subgraphs can be enumerated by frequent subgraph mining algorithms
- **Proposition:**
  The set of testable subgraphs $\tau(\mathcal{H})$ coincides with the set of frequent subgraphs with the threshold σ_{rt} s.t.

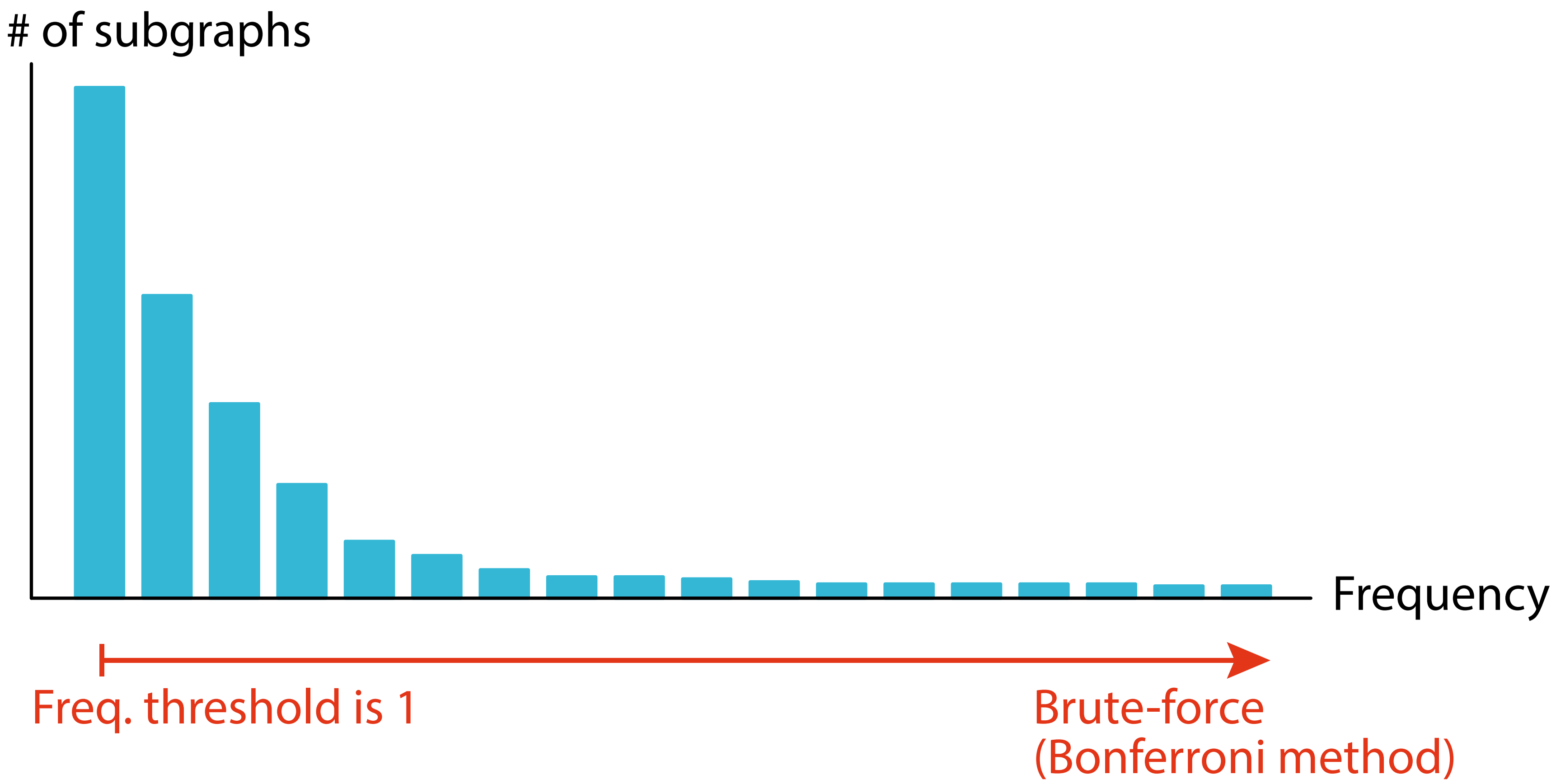
  $$\#\text{ of subgraphs with minfreq } \sigma_{\mathrm{rt}} - 1 > \alpha / \psi(\sigma_{\mathrm{rt}} - 1),$$
  $$\#\text{ of subgraphs with minfreq } \sigma_{\mathrm{rt}} \le \alpha / \psi(\sigma_{\mathrm{rt}}),$$

  - $\alpha / \psi(\sigma)$ shows the admissible number of subgraphs at σ
    - $\psi(\sigma) = \binom{n}{\sigma} / \binom{n+n'}{\sigma}$ (Minimum P value at σ)
    - For $k_{\mathrm{rt}} = \alpha / \psi(\sigma_{\mathrm{rt}})$, if ψ is monotonically decreasing, $m(k_{\mathrm{rt}}) = |\{ H \in \mathcal{H} \mid \psi(f(H)) \le \psi(\sigma_{\mathrm{rt}}) \}| = |\{ H \in \mathcal{H} \mid f(H) \ge \sigma_{\mathrm{rt}} \}|$

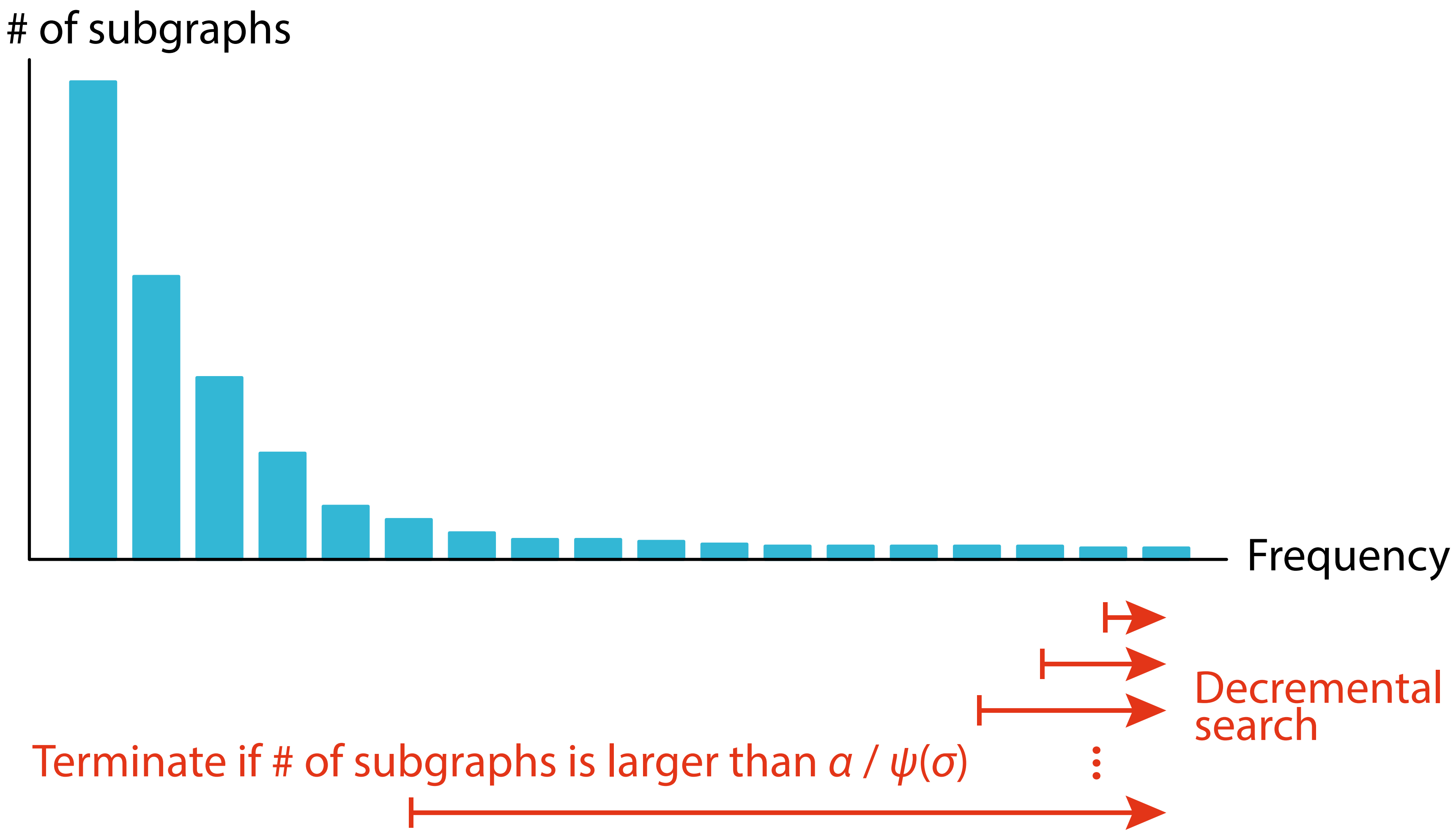
How to Use Subgraph Mining



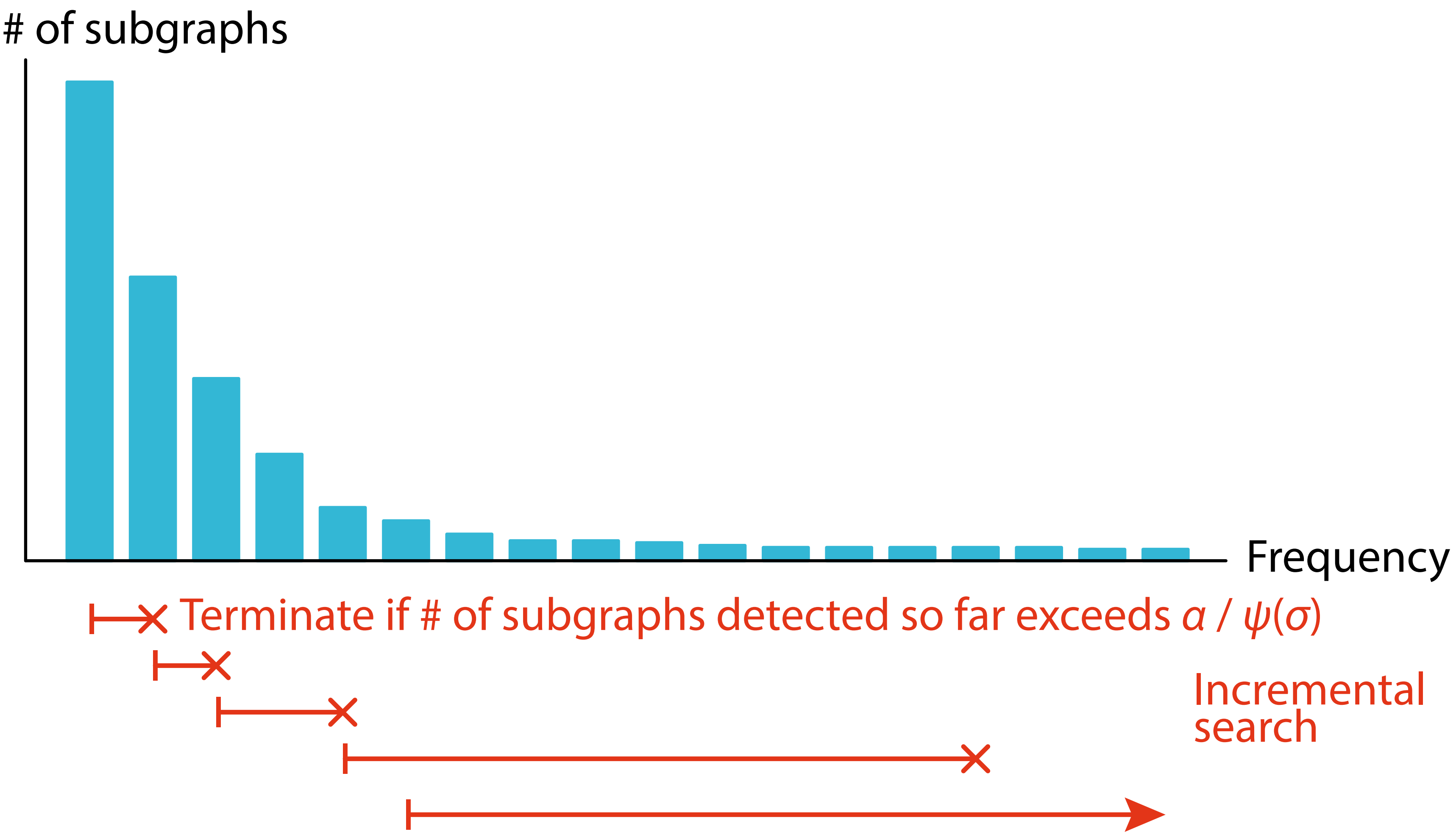
Brute-Force Search (Bonferroni)



Decremental Search (LAMP)



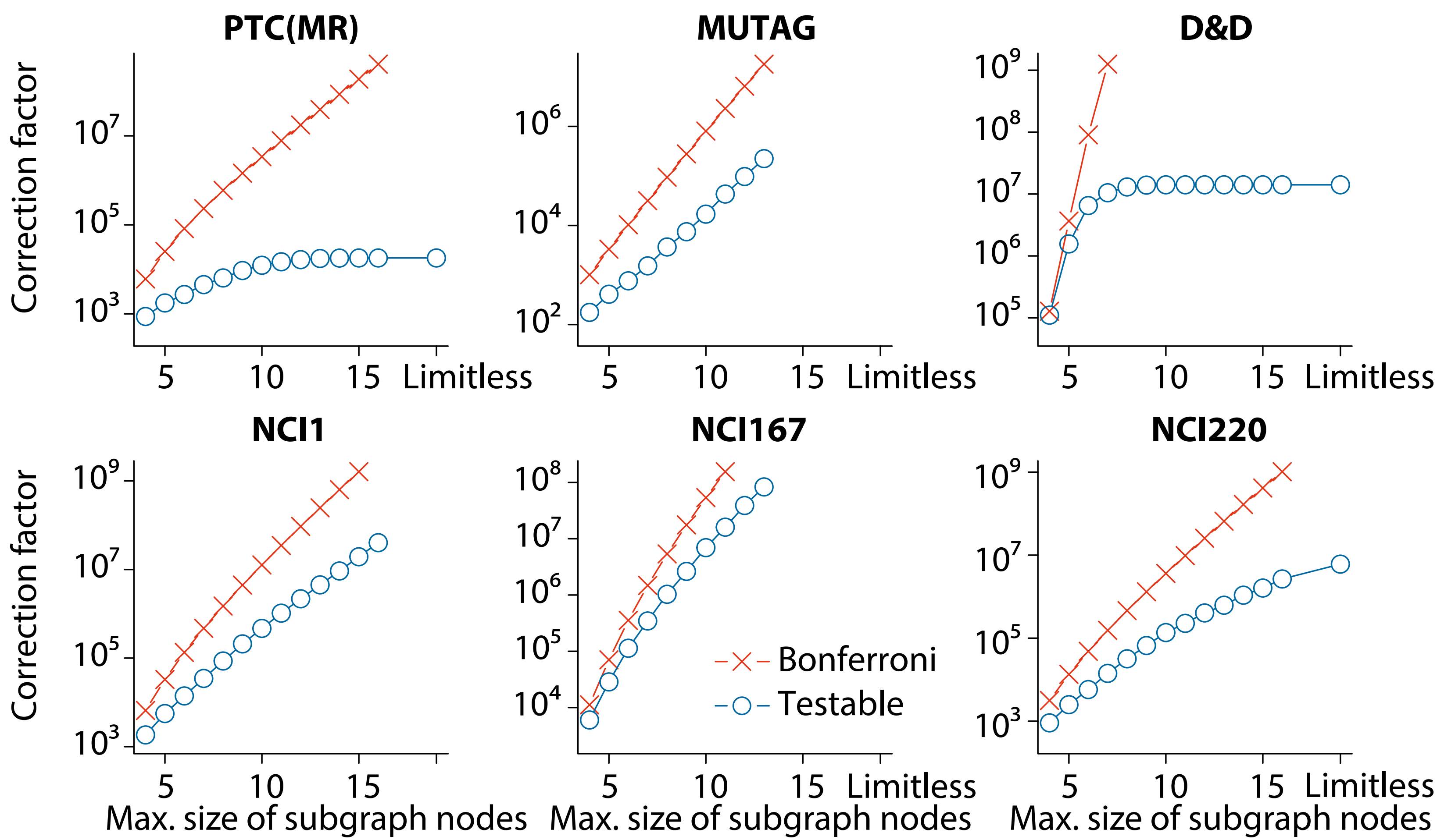
Incremental Search



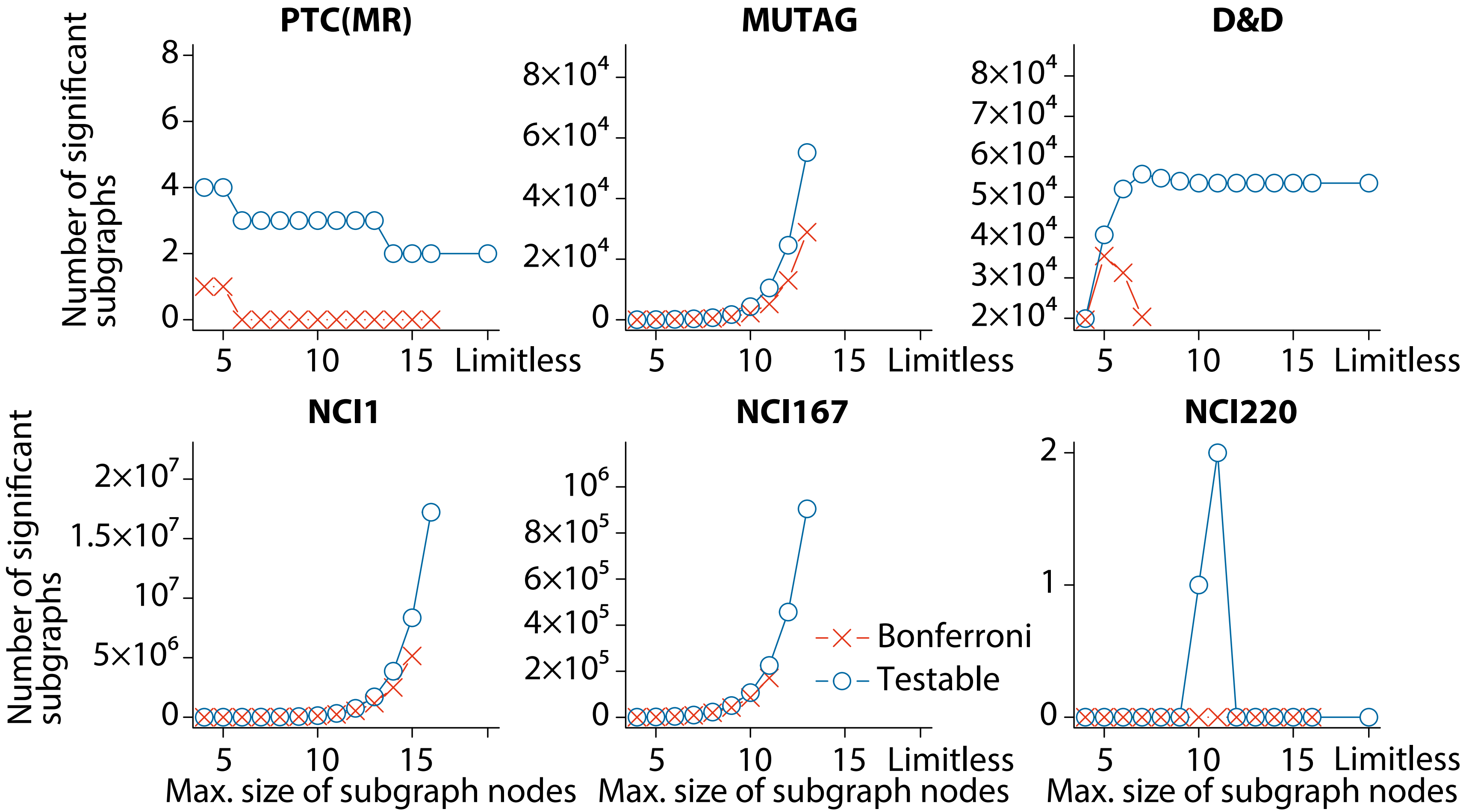
Datasets

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

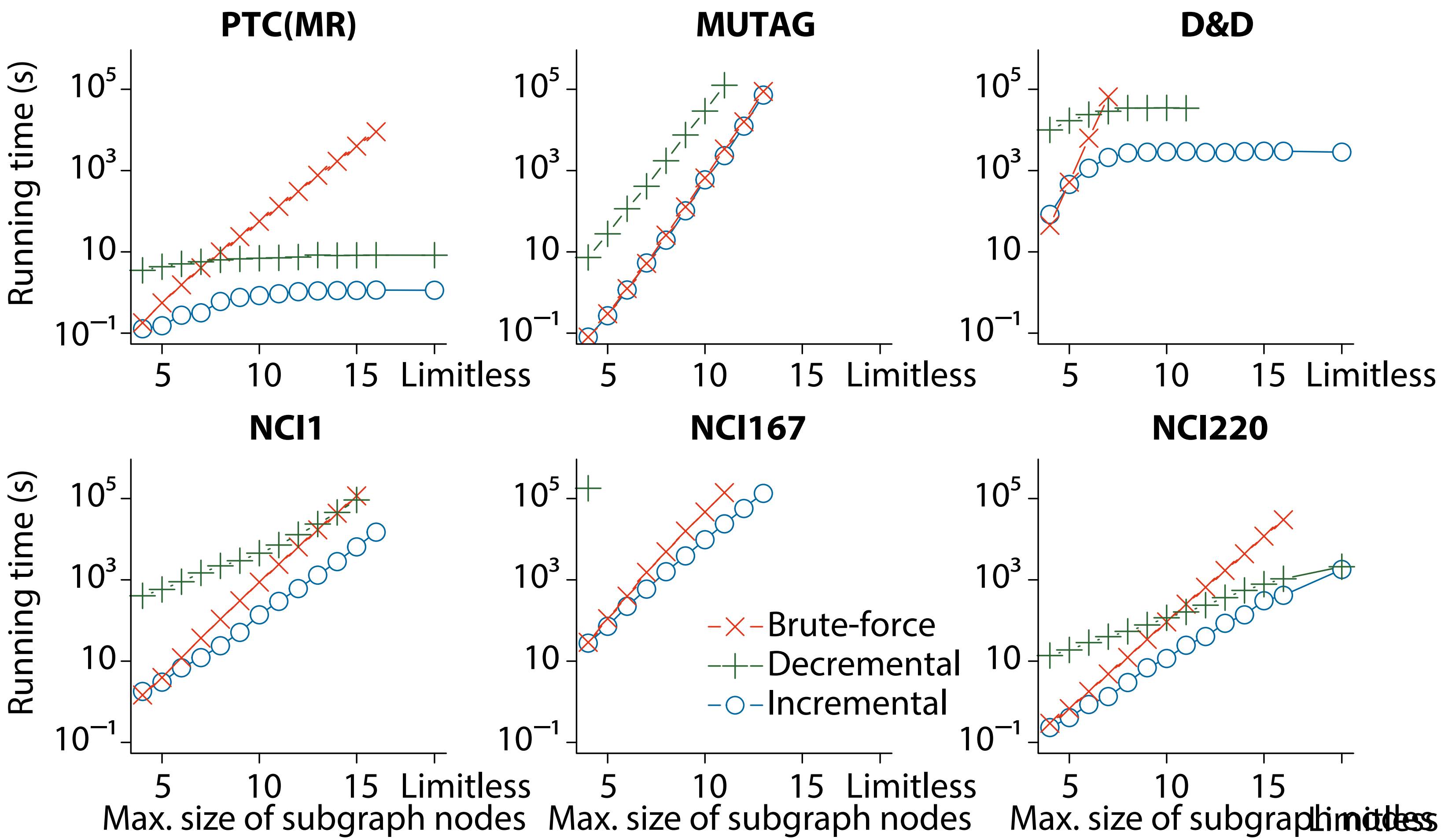
Correction Factor



Number of Significant Subgraphs



Running Time (second)



Running Time Summary

- RMSD (root mean square deviation) of running time (seconds) to the best (fastest) running time on all datasets

Brute-force	Decremental (LAMP)	Incremental
6.994×10^4	2.410×10^4	1.230×10^2

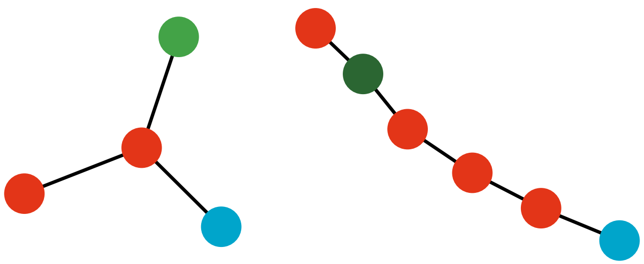
- Incremental search is the fastest
  - More than two orders of magnitude faster than brute-force
  - Much faster than decremental (LAMP) as the final minimum support is usually small (~20)

Final Minimum Supports

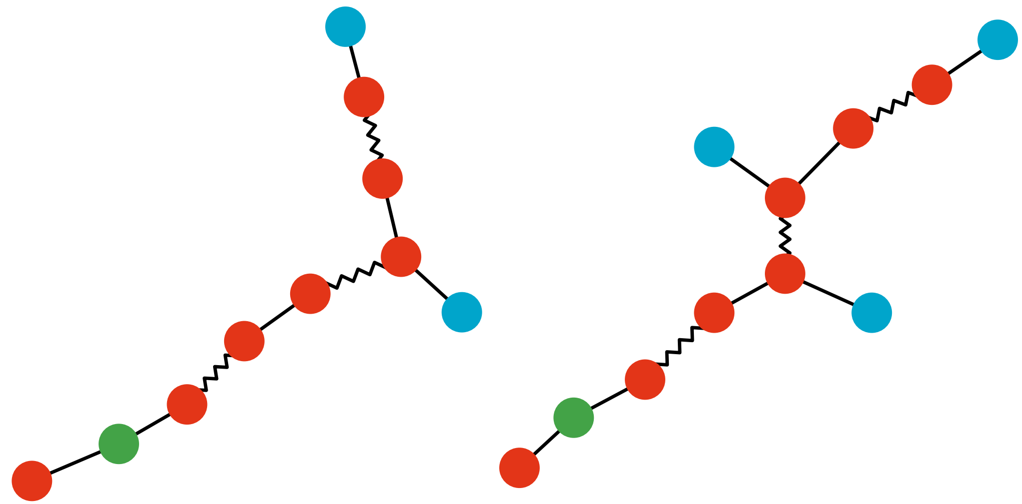
Dataset	Maximum size of subgraph nodes							n
	5	7	9	11	13	15	Limitless	
PTC(MR)	9	10	11	11	11	11	11	181
MUTAG	8	10	11	12	14	—	—	125
D&D	20	22	22	22	22	22	22	691
NCI1	17	20	22	25	27	29	—	2104
NCI167	7	8	9	10	11	—	—	9615
NCI220	10	11	13	14	15	16	18	290

Detected Significant Subgraphs

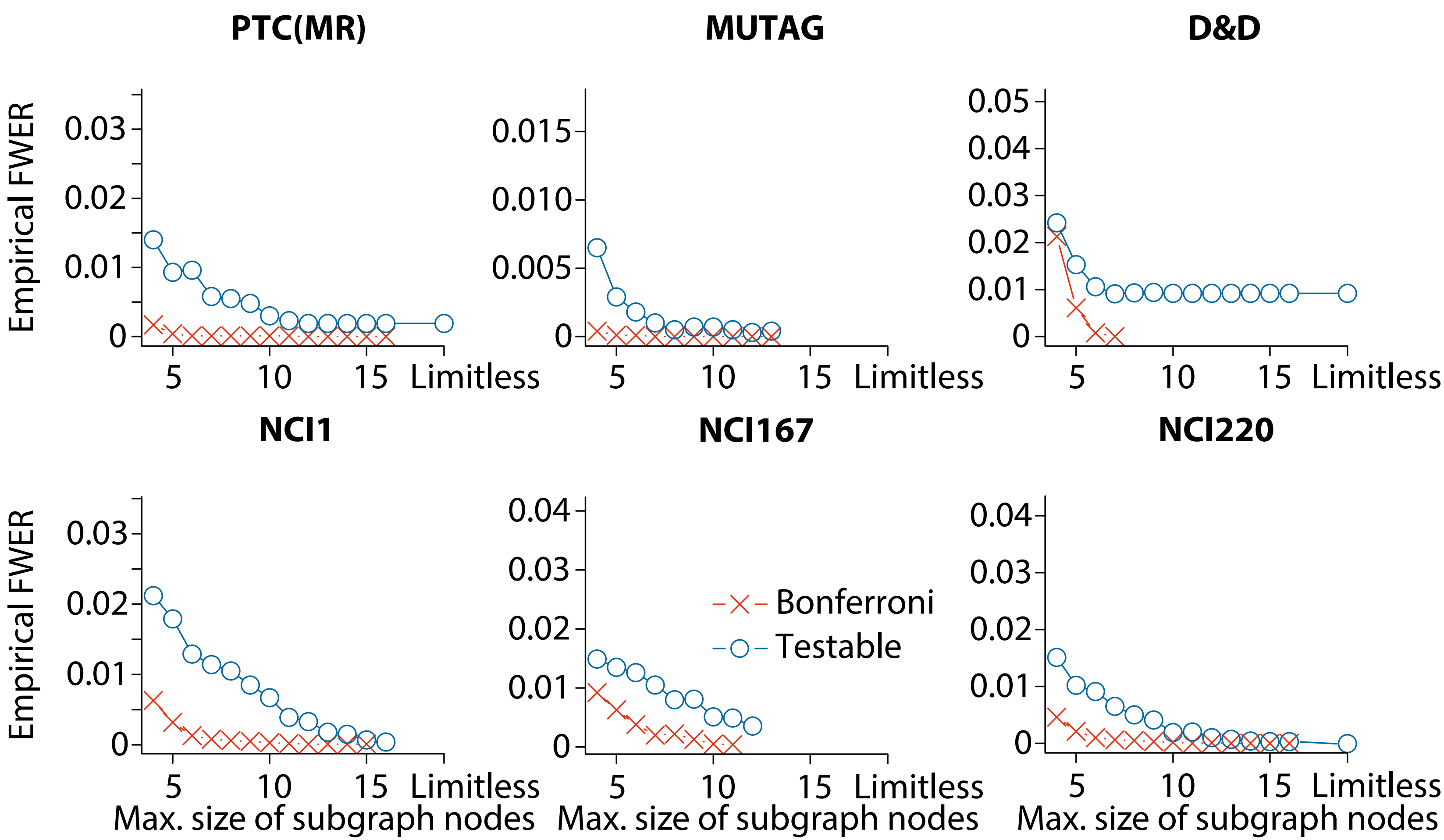
PTC (MR)
(carcinogenicity)



NCI 220
(anti-cancer activity)



FWER Is still Too Low!



Related work: LAMP version 2

- Minato et al. proposed a faster version of LAMP in itemset mining
  - Minato, S., Uno, T., Tsuda, K., Terada, A. and Sese, J.:
    Fast Statistical Assessment for Combinatorial Hypotheses Based on Frequent Itemset Mining
    ECML PKDD 2014
- The idea is almost the same with our incremental search
  - Start from $\sigma = 1$, every time an item is added, the condition $|\mathcal{I}(\sigma)| \leq \alpha/\psi(\sigma)$ is checked
    - $\mathcal{I}(\sigma)$: the set of itemsets found so far with the frequency $\geq \sigma$
  - As soon as $|\mathcal{I}(\sigma)| > \alpha/\psi(\sigma)$, the current σ is too large and we decrement it

Conclusion

- Significant subgraphs mining with multiple testing correction is achieved
  - The first work that considers multiple testing correction in graph mining
- Efficient and effective (less false negatives) using testability
- Future work
  - Increase the FWER with keeping $\leq \alpha$
    - Currently we ignore correlations between subgraphs

Papers about Testability

- Tarone, R.E.:
  A modified Bonferroni method for discrete data
  Biometrics (1990)

- Terada, A., Okada-Hatakeyama, M., Tsuda, K., Sese, J.:
  Statistical significance of combinatorial regulations,
  Proc. Natl. Acad. Sci. USA (2013).

- Minato, S., Uno, T., Tsuda, K., Terada, A., Sese, J.:
  Fast Statistical Assessment for Combinatorial Hypotheses Based on Frequent Itemset Mining
  ECML PKDD 2014

- Sugiyama, M., Llinares López, F., Kasenburg, N., Borgwardt, K.M.:
  Significant Subgraph Mining with Multiple Testing Correction,
  SIAM SDM 2015 (`http://arxiv.org/abs/1407.0316`)
  - Code: `http://git.io/N126`