April 30, 2015
SDM15

# Significant Subgraph Mining with Multiple Testing Correction


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)

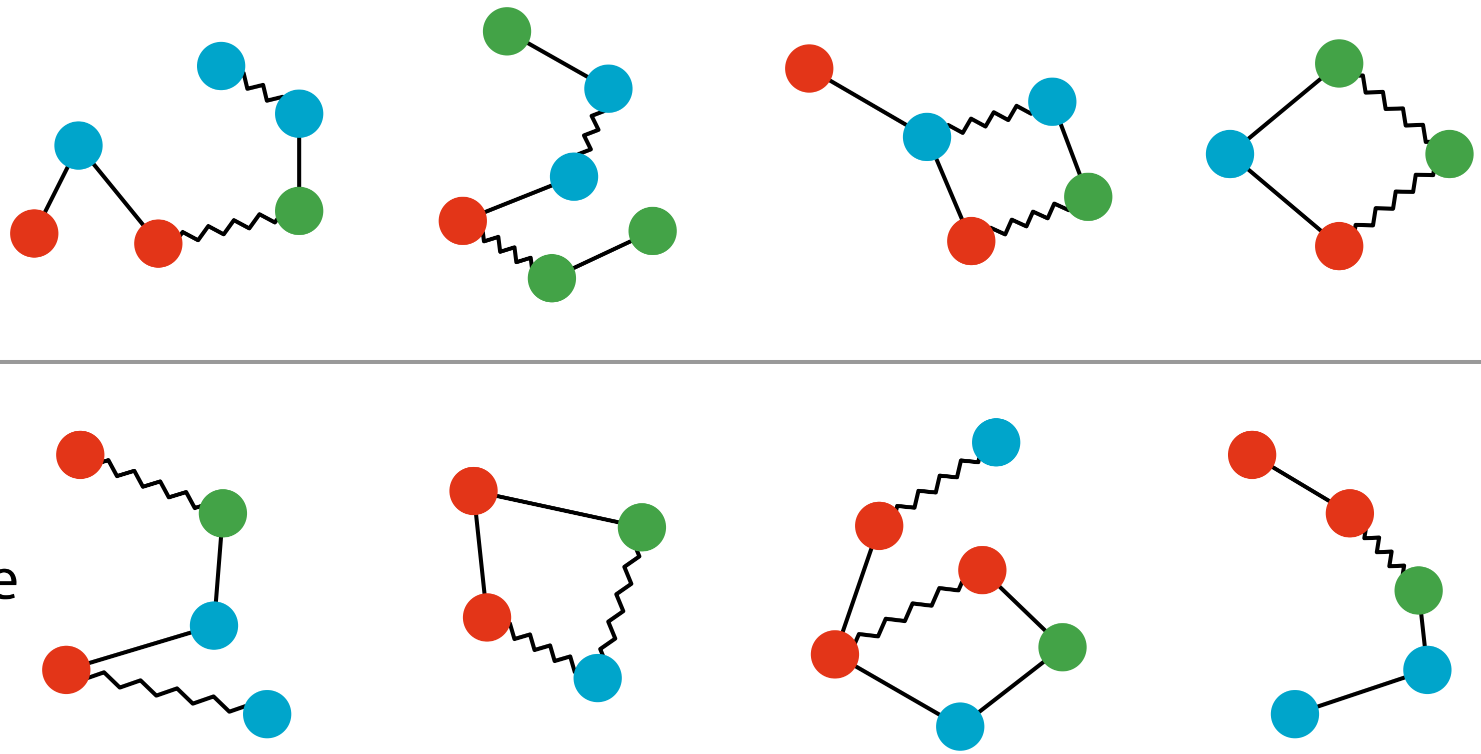
# Summary

- **Problem:** Given a collection of graphs with class labels, find all subgraphs whose occurrences are significantly enriched in a particular class
  - A central step for deep understanding
- **Difficulty:** The number of subgraphs is massive (often more than a billion!)
  - Computationally expensive
  - Need of multiple testing correction to control false positive rate
- **Solution:** Only examining testable subgraphs
  - The number of candidate subgraphs dramatically reduced
  - Rigorous multiple testing correction

# Find Associated Subgraphs

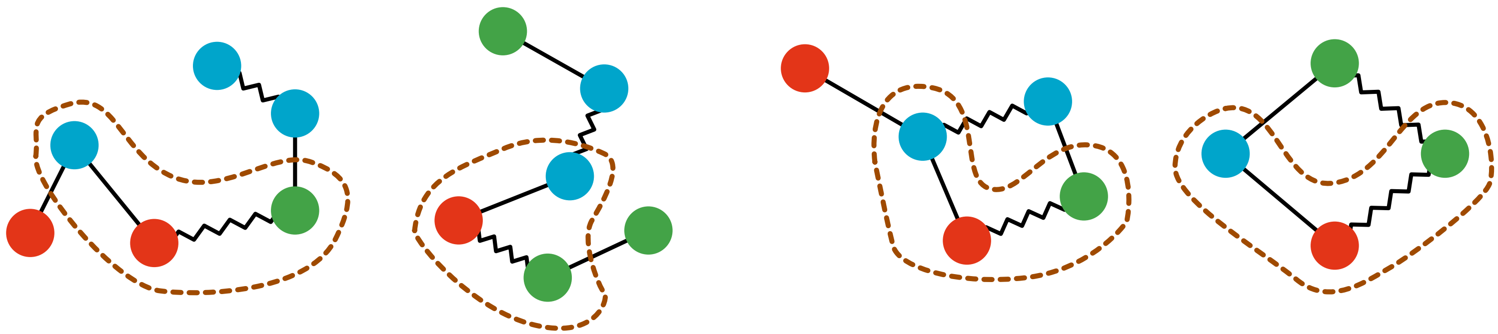
Active

Inactive

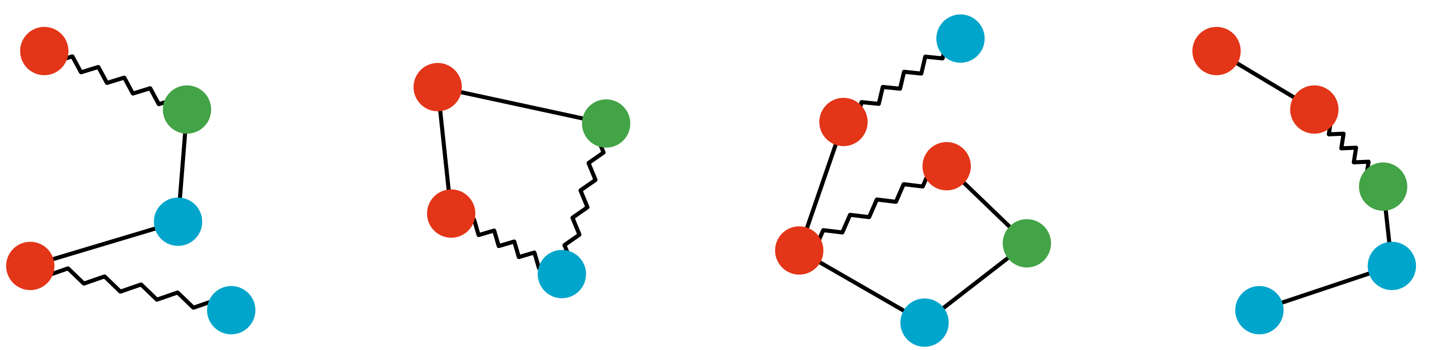


# Find Associated Subgraphs

Active



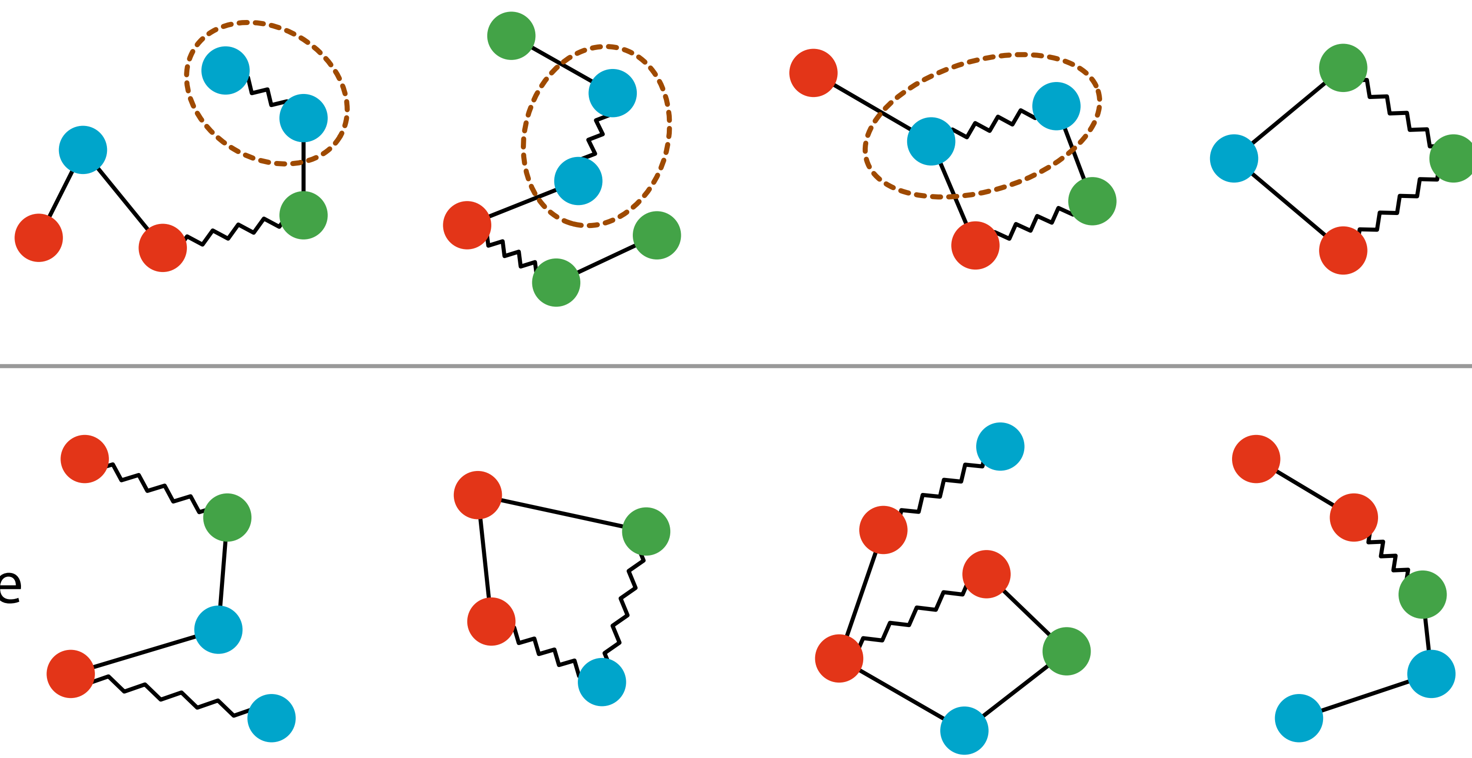
Inactive



# Find Associated Subgraphs

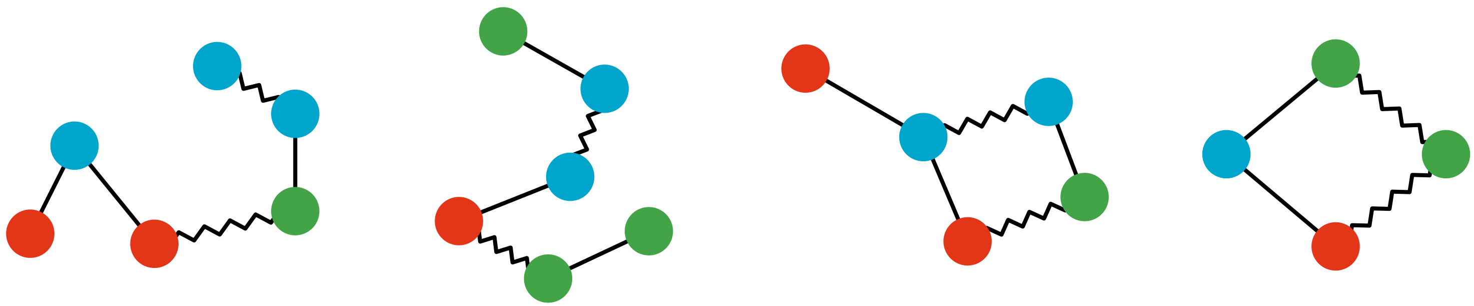
Active

Inactive

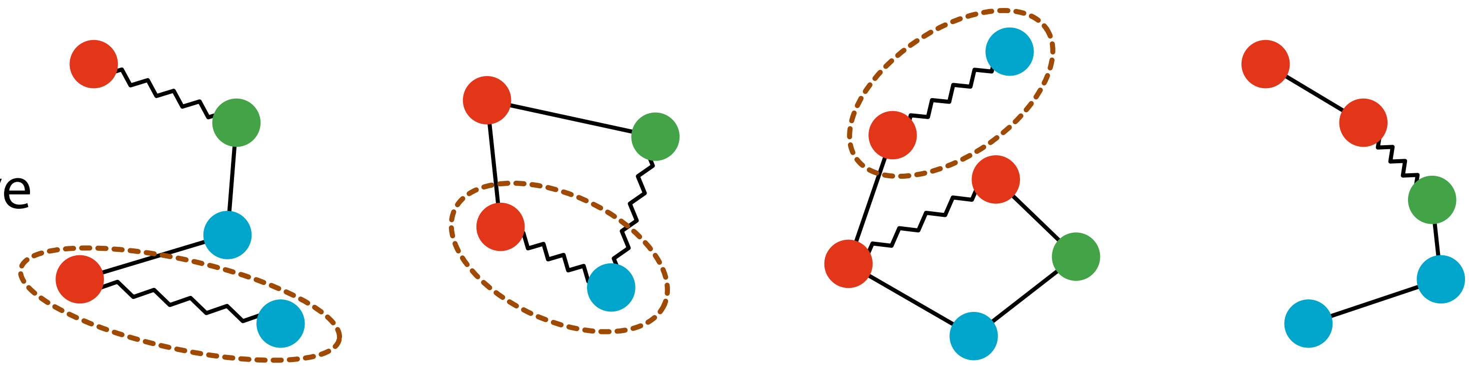


# Find Associated Subgraphs

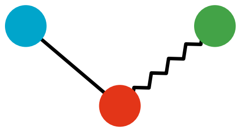
Active



Inactive



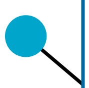
# Multiple Testing



| | 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



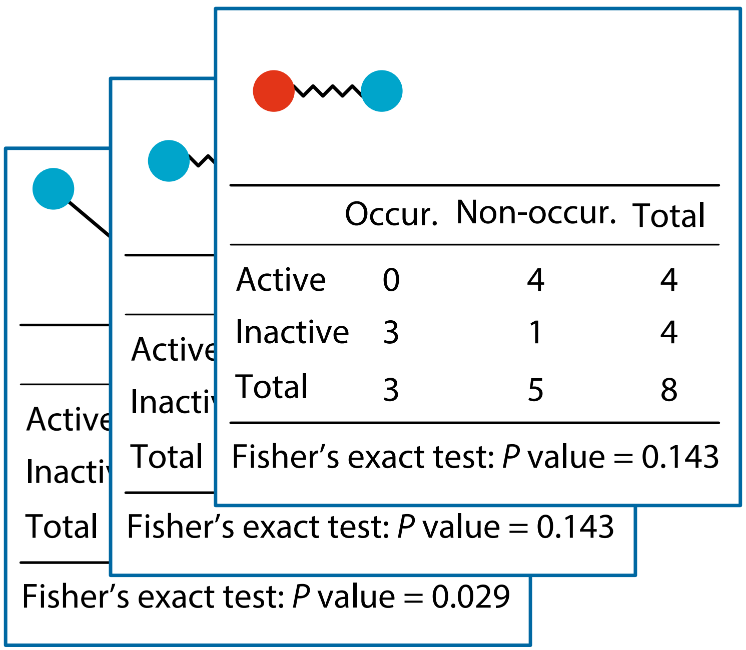
Active
Inacti
Total

Fisher's exact test: $P$ value = 0.029

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

Fisher's exact test: $P$ value = 0.143

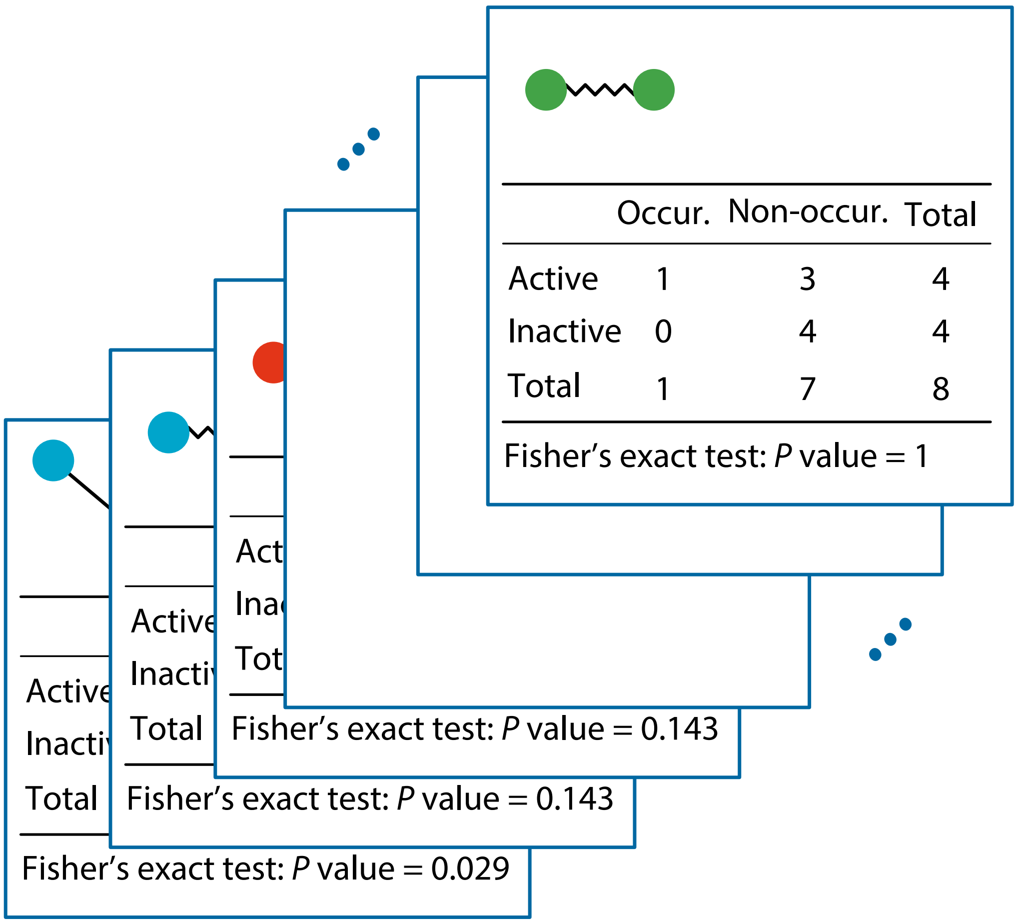
# Multiple Testing



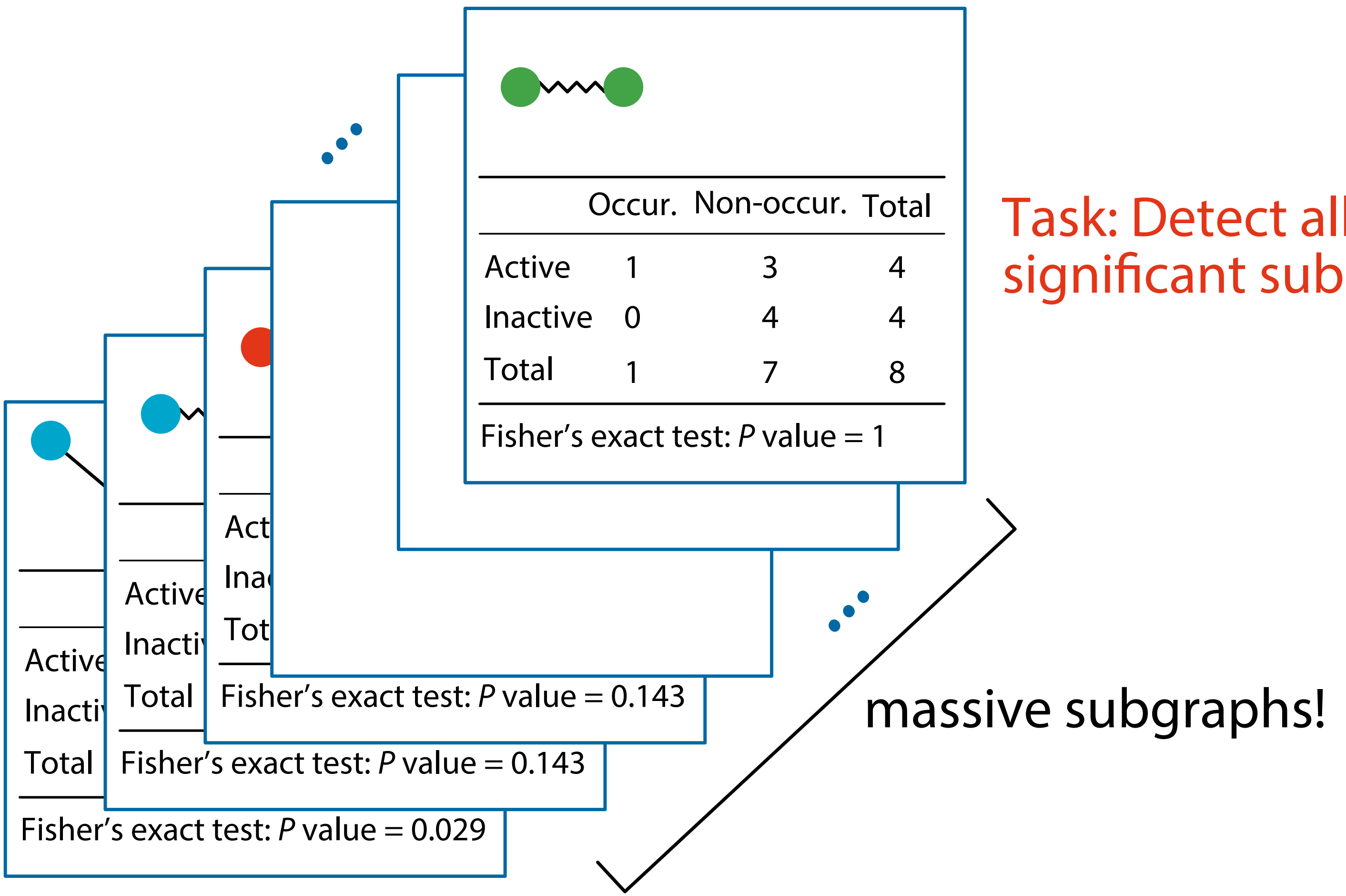

| | Occur. | Non-occur. | Total |
|---|---|---|---|
| Active | 0 | 4 | 4 |
| Inactive | 3 | 1 | 4 |
| Total | 3 | 5 | 8 |

Fisher's exact test: $P$ value = 0.143

# Multiple Testing



# Multiple Testing



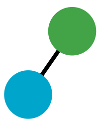

Task: Detect all significant subgraphs

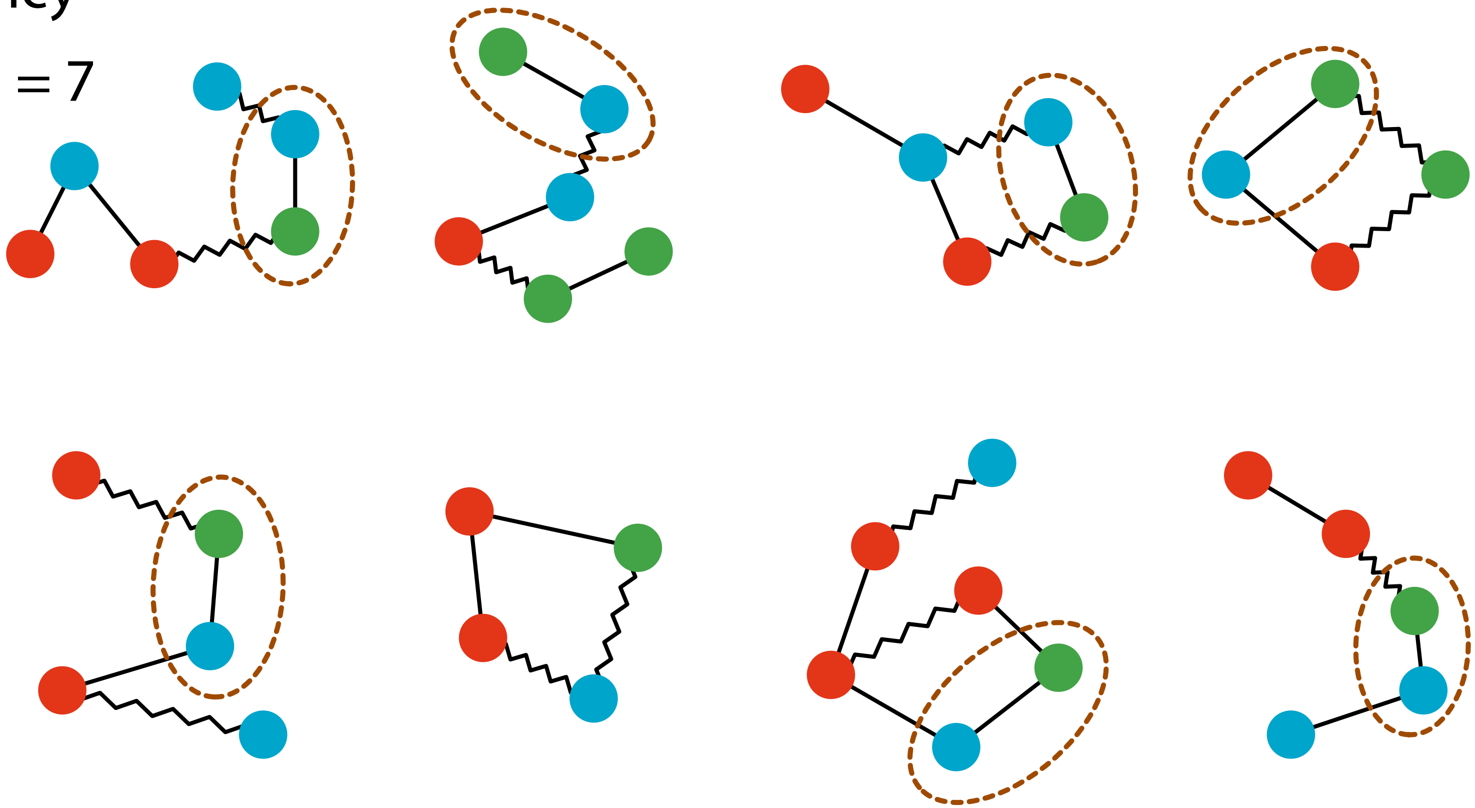
# Multiple Testing Correction

- If we test $m$ subgraphs, $\alpha m$ subgraphs are false positives
  - $\alpha$: Significance level (predetermined by the user)
- FWER: Probability of having more than one false positives among all subgraphs
  - FWER = Pr(FP > 0)
    - FP: Number of false positives
- To achieve FWER = $\alpha$, change the significance level for each test from $\alpha$ to $\delta$
  - $\delta$: corrected significance level
  - $\delta \leq \alpha$
    - Bonferroni correction is popular: $\delta^*_{\mathrm{Bon}} = \alpha / m$

# Counting the Frequency of Subgraphs

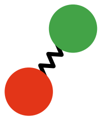
Frequency

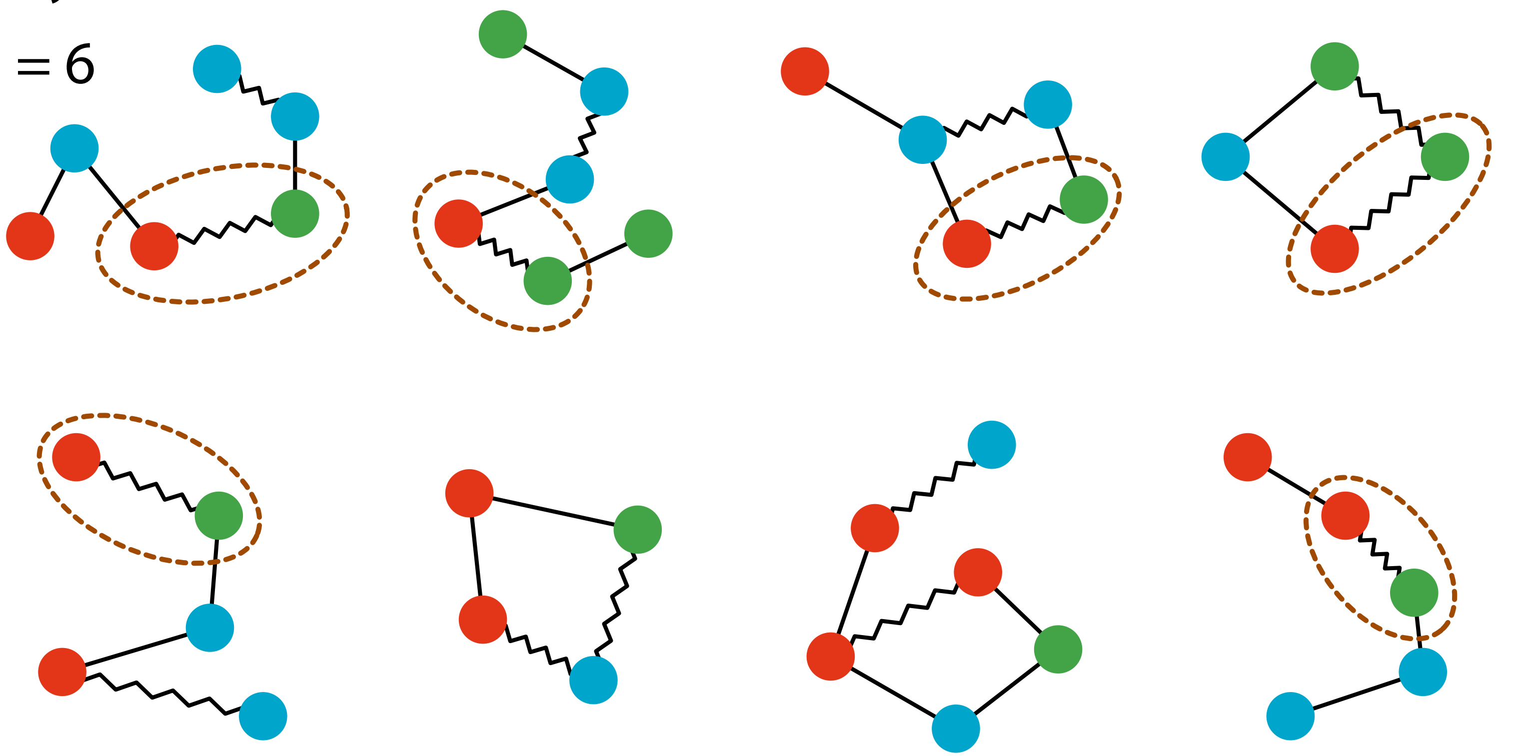
$f($  $) = 7$



# Counting the Frequency of Subgraphs

Frequency

*f* (  ) = 6



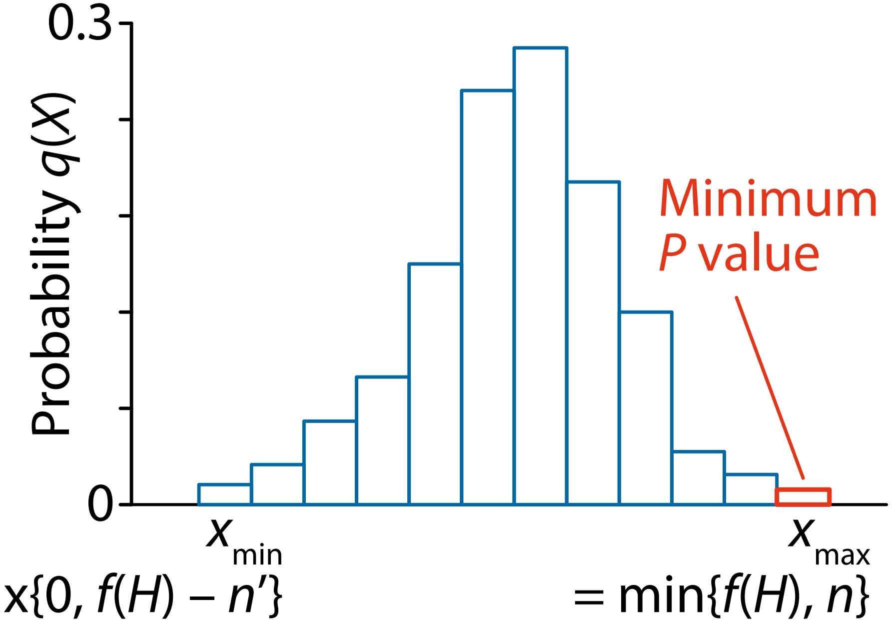
# The Minimum $P$ Value

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

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

| | $\boxed{\text{Occ.}}$ | 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
is determined from the frequency $f(H)$ of a subgraph $H$:

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

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

  *For a hypothesis H, if its minimum P value is larger 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, $\alpha/k$: 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 is to optimize $k$:

$$k^* = \operatorname*{argmax}_k m(k) \quad \text{s.t. } m(k) \le k$$

  - Enumerate testable subgraphs whose min. $P$ values < $\alpha/k^*$

$$\delta^*_{\mathrm{Bon}} = \alpha/(\text{\# of all subgraphs})$$

$$\delta^*_{\mathrm{Tar}} = \alpha/(\text{\# of testable subgraphs})$$

# Subgraphs Are Testable Iff Frequent

- Our task:

$$k^* = \operatorname*{argmax}_k m(k) \quad \text{s.t. } m(k) \le k$$

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

# Subgraphs Are Testable Iff Frequent

- Our task:

$$k^* = \operatorname*{argmax}_k m(k) \quad \text{s.t. } m(k) \le k$$

$$\Downarrow$$

$$\sigma^* = \operatorname*{argmax}_\sigma m'(\sigma) \quad \text{s.t. } m'(\sigma) \le \alpha/\psi(\sigma)$$

  - $m(k)$ = # of subgraphs whose minimum $P$ values $< \alpha/k$
  - $m'(\sigma)$: # of subgraphs whose frequency $\ge \sigma$
    - # of “frequent subgraphs”
  - $\psi(\sigma)$: the minimum $P$ value at $\sigma$, $\psi(\sigma) = \binom{n}{\sigma} / \binom{n+n'}{\sigma}$

# Subgraphs Are Testable Iff Frequent

- Our task:

$$k^* = \operatorname*{argmax}_k m(k) \quad \text{s.t. } m(k) \le k$$

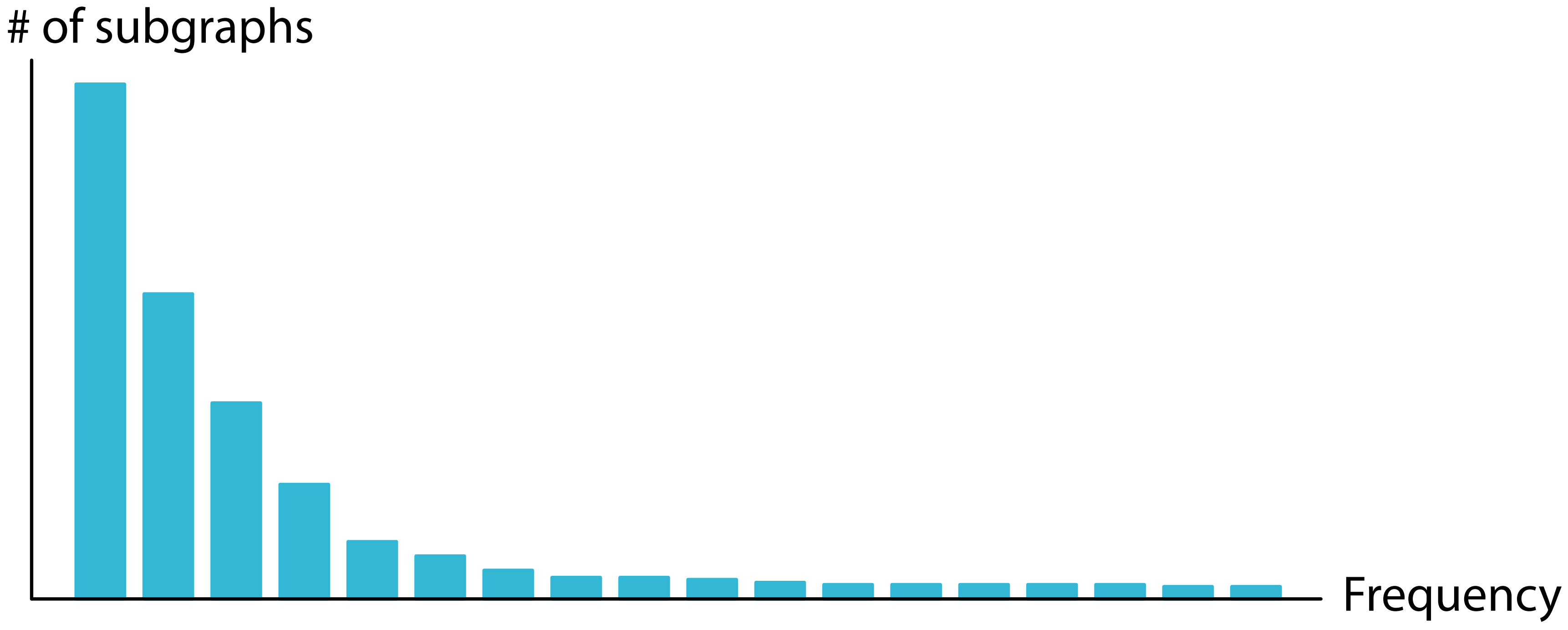
$$\Downarrow$$

$$\sigma^* = \operatorname*{argmax}_\sigma m'(\sigma) \quad \text{s.t. } m'(\sigma) \le \alpha / \psi(\sigma)$$

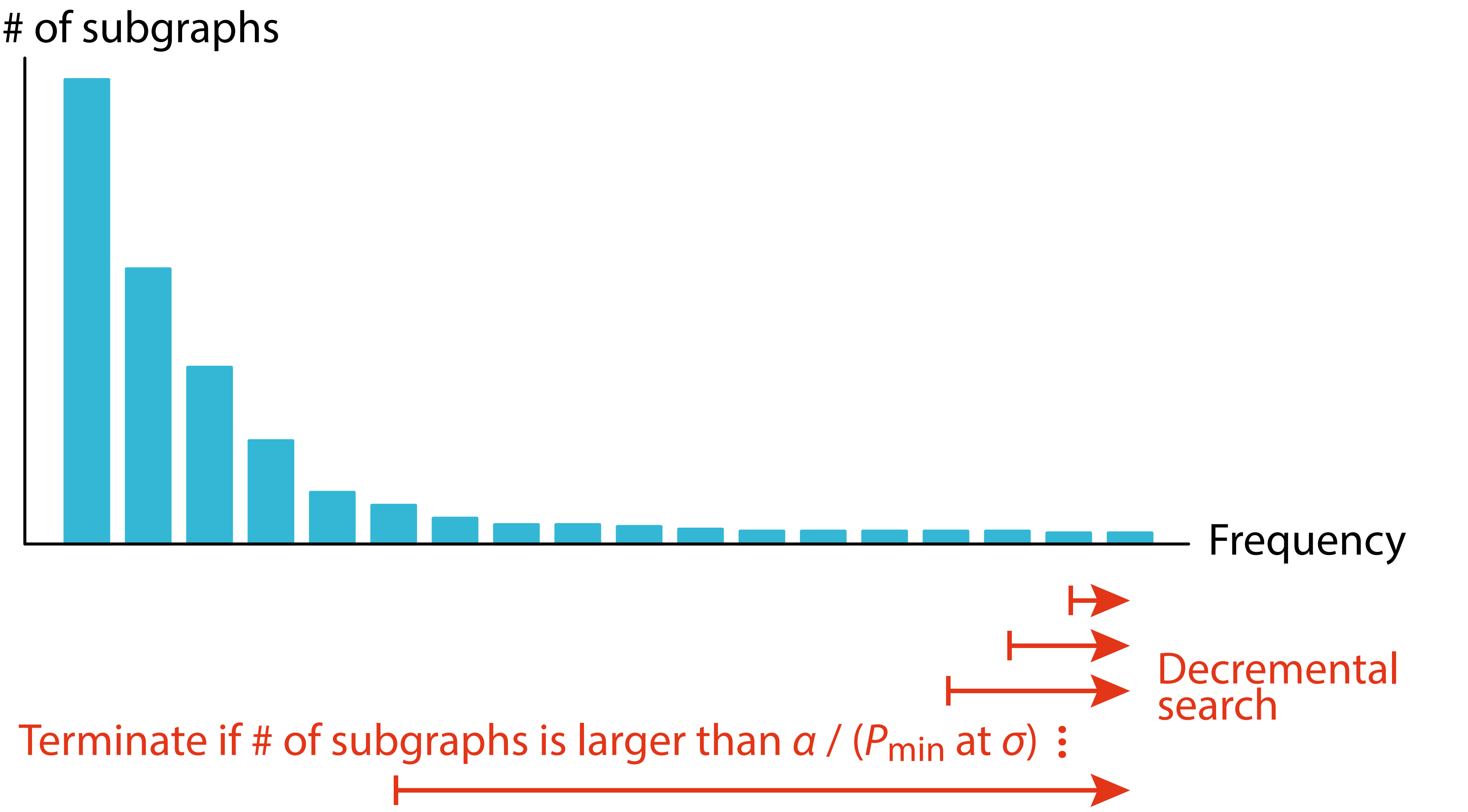
  - $m(k)$ = # of subgraphs whose minimum $P$ values $< \alpha / k$
  - $m'(\sigma)$: # of subgraphs whose frequency $\ge \sigma$
    - # of "frequent subgraphs"
  - $\psi(\sigma)$: the minimum $P$ value at $\sigma$, $\psi(\sigma) = \binom{n}{\sigma} / \binom{n+n'}{\sigma}$

Testable subgraphs = Frequent subgraphs

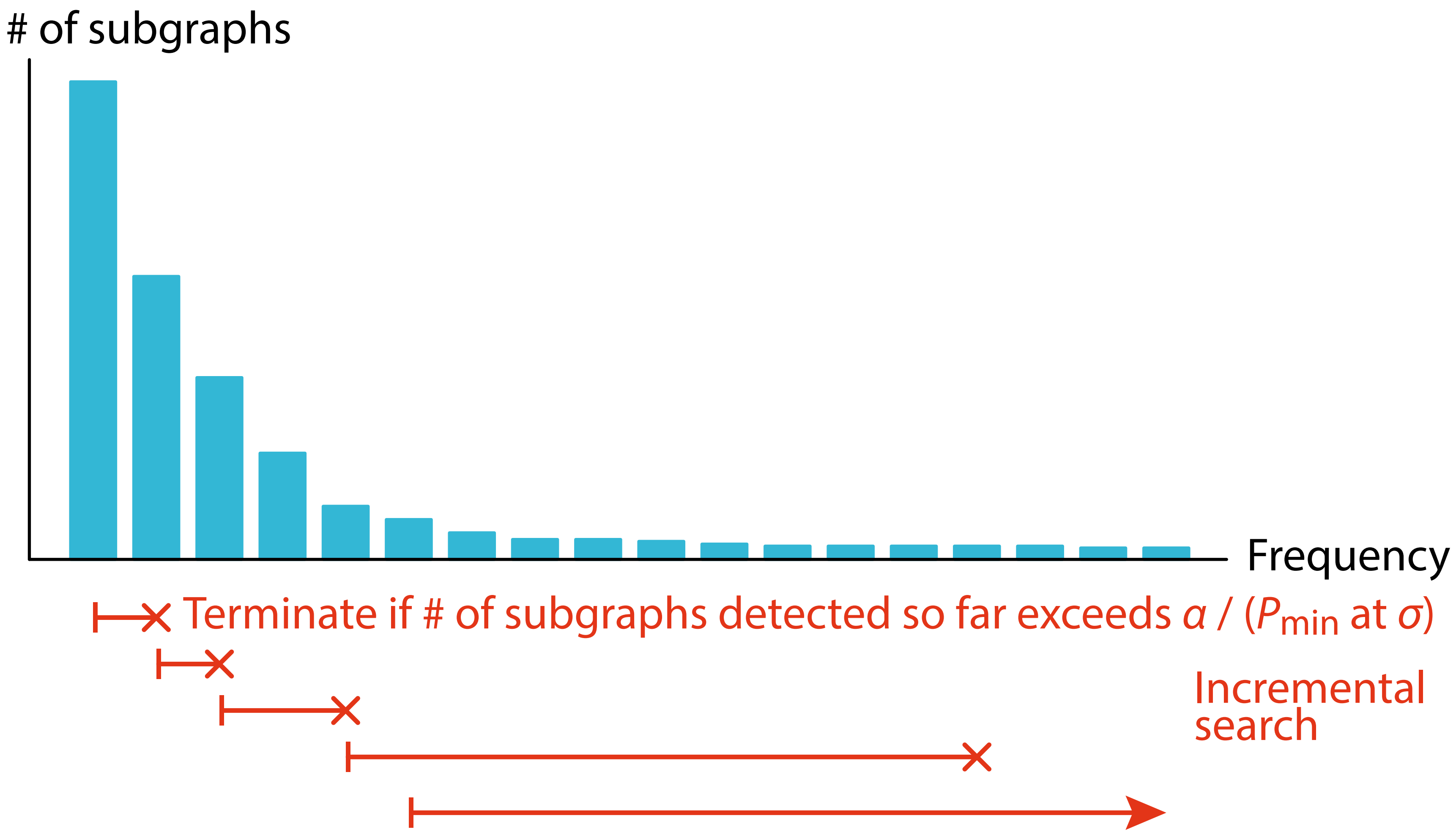
# How to Use Subgraph Mining



# Decremental Search (LAMP)



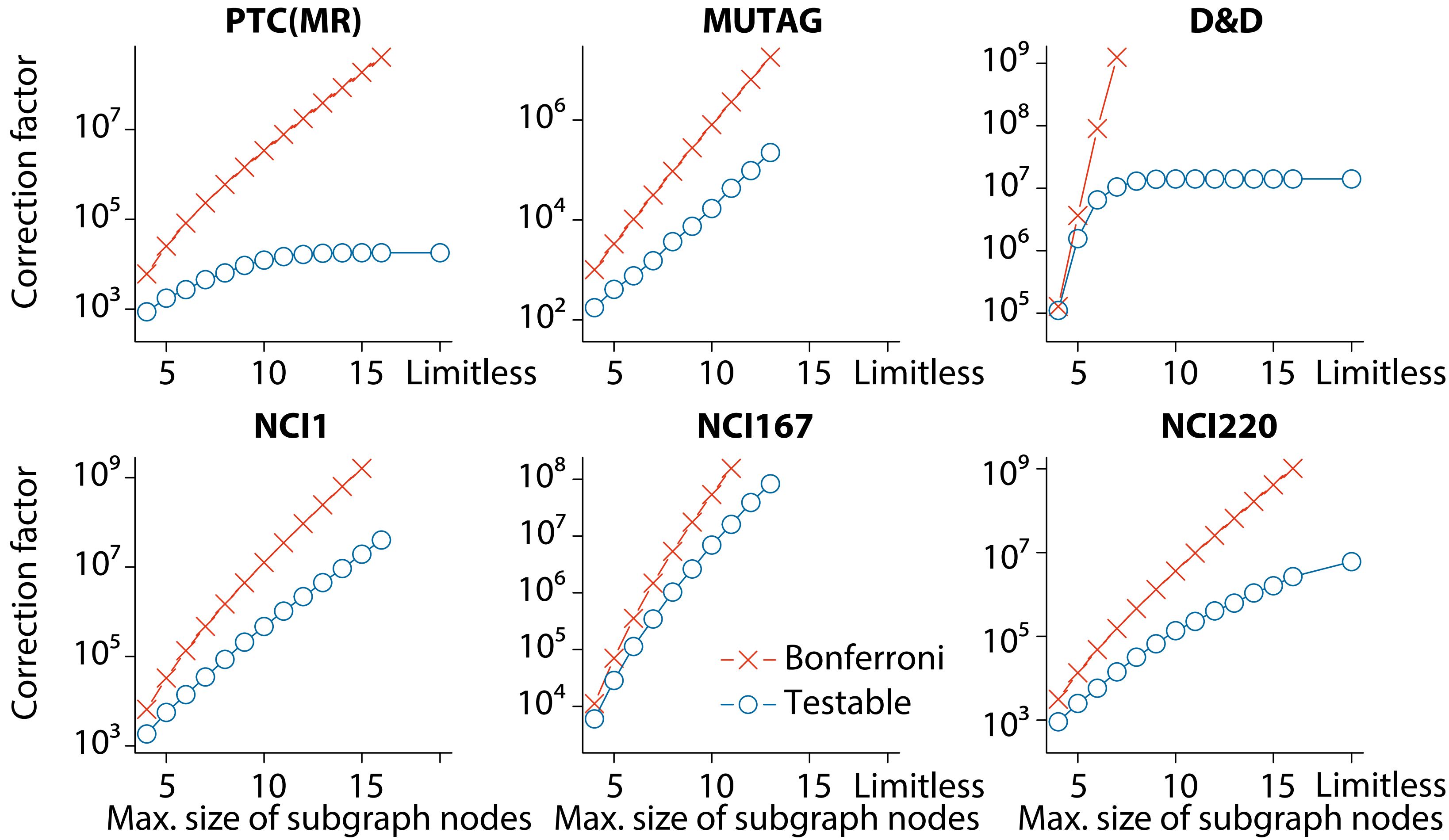
# Incremental Search



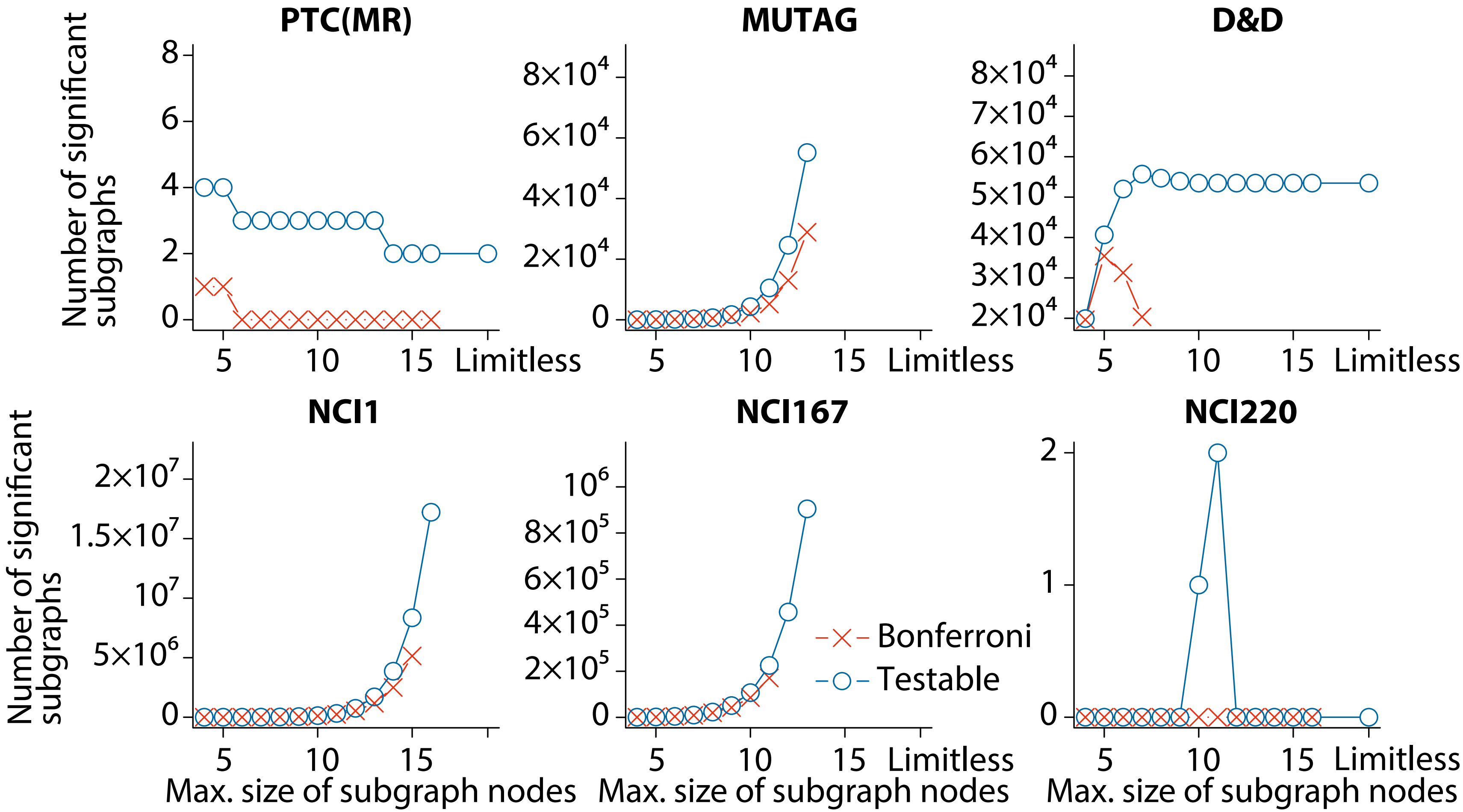
# Datasets

| 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 |

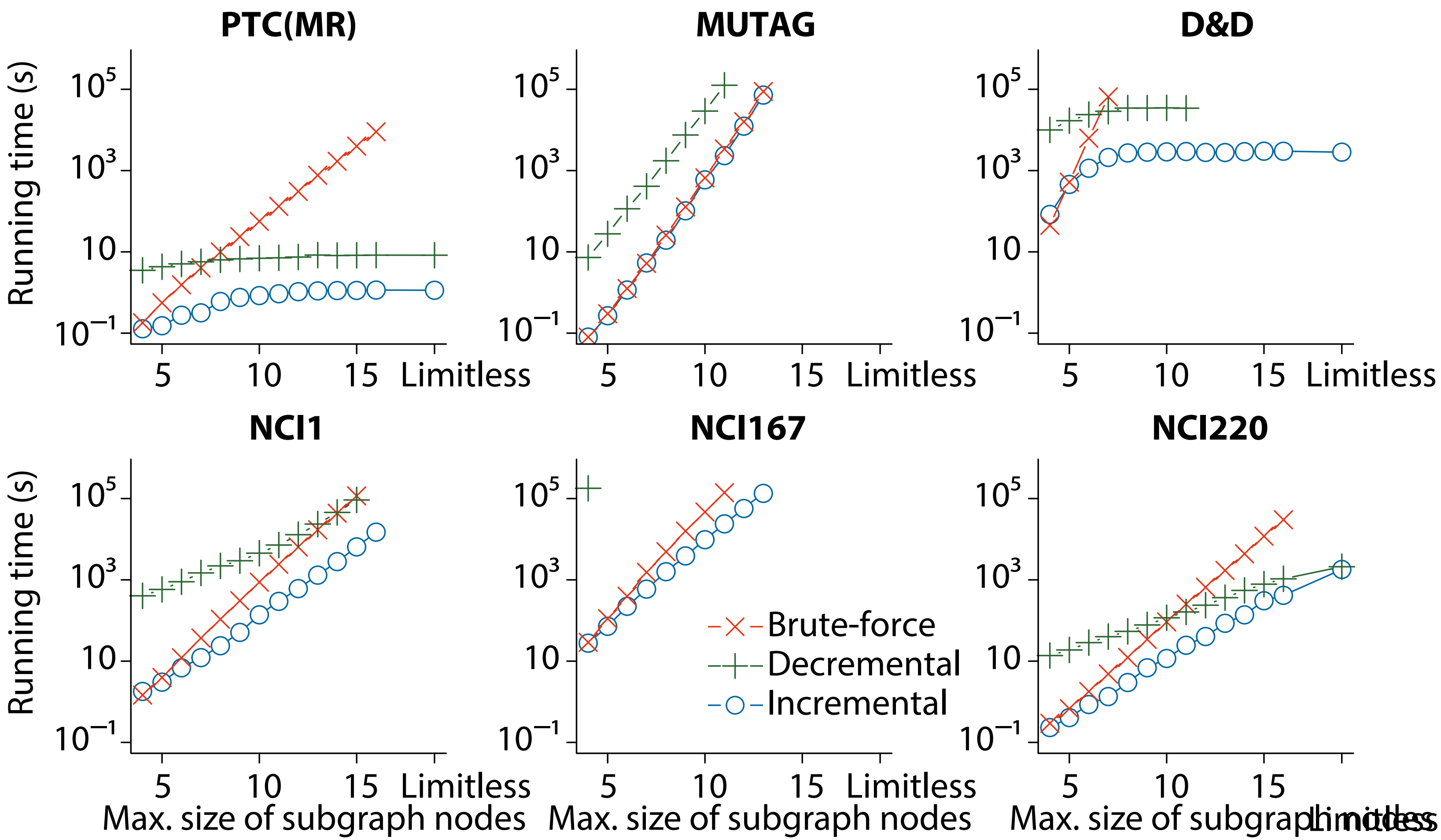
# Correction Factor



# 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 |
|---|---|---|
| $6.994 \times 10^4$ | $2.410 \times 10^4$ | $1.230 \times 10^2$ |

- 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 Frequencies

| 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 |

# Conclusion

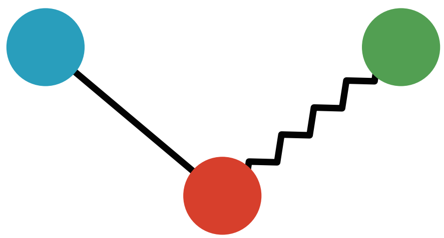
- We achieved to enumerate all significant subgraphs
  - The first work that considers multiple testing correction in graph mining
- Efficient and more powerful (less false negatives) using testability and frequent subgraph mining
- Pattern mining, a classical yet central topic in data mining, can be enriched by introducing statistical assessment
  - Can be applied in scientific fields such as biology

# Appendix

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

# 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 |

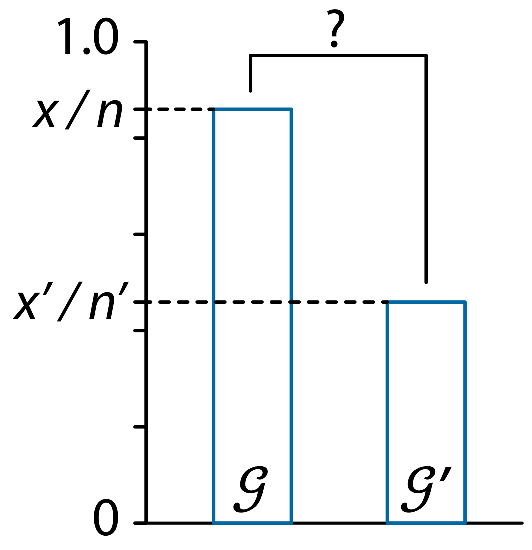
**Null hypothesis:** The occurence of the subgraph is independent from the activity

**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 $\mathcal{G}'$
  - $|\mathcal{G}| = n$, $|\mathcal{G}'| = n'$ $(n \le n')$
- 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 | $x + x'$ | $(n - x) + (n' - x')$ | $n + n'$ |

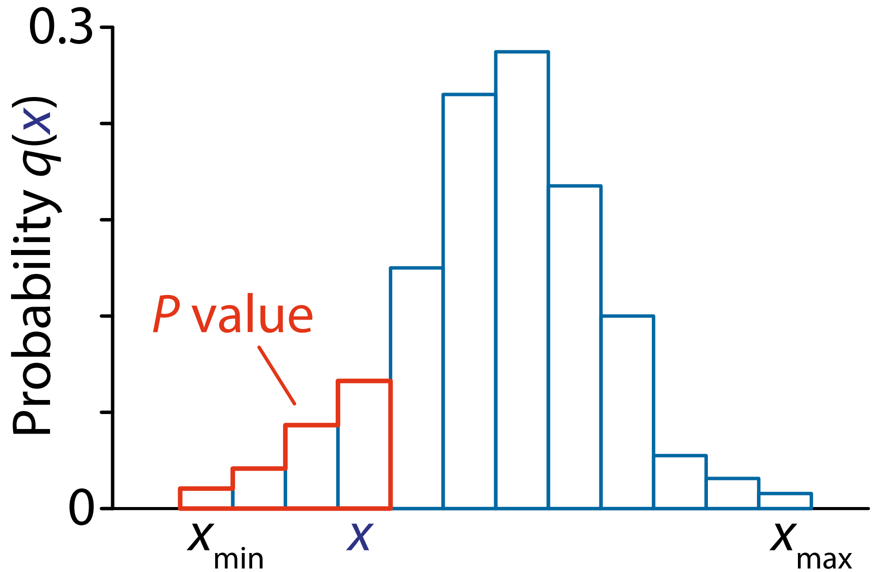


# Fisher's Exact Test

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

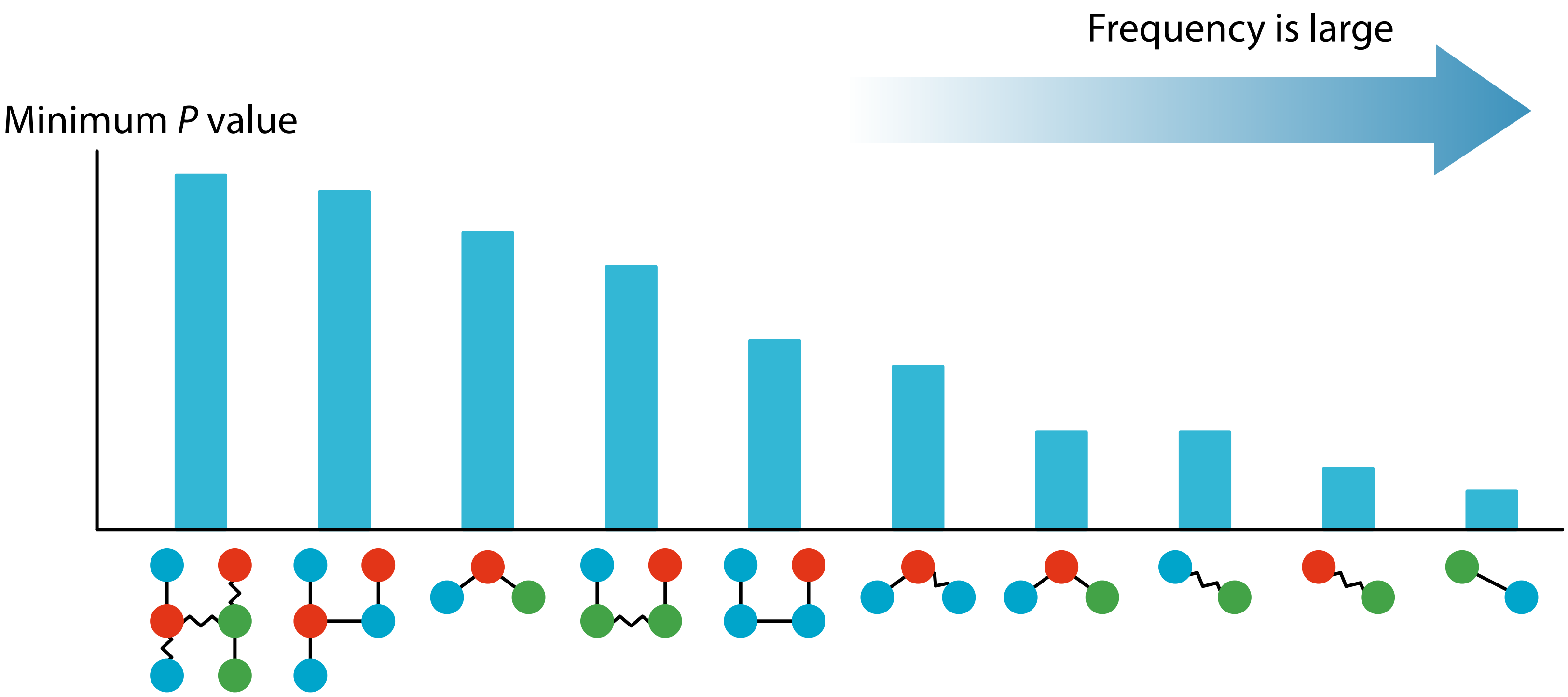
$$q(x) = \binom{n}{x}\binom{n'}{x'} \Big/ \binom{n+n'}{x+x'}$$

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

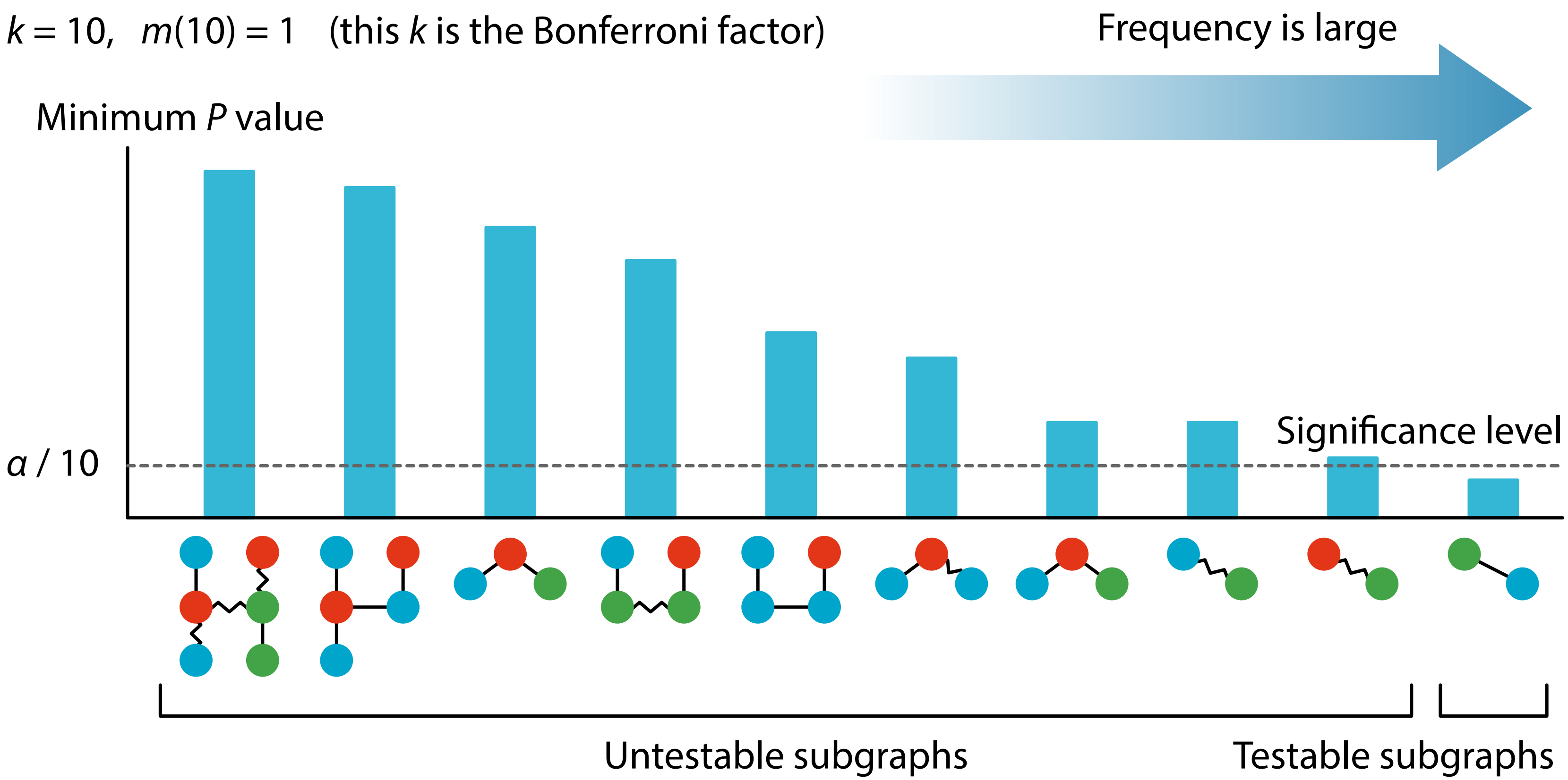



$= \max\{0, x + x' - n'\}$ $= \min\{x + x', n\}$

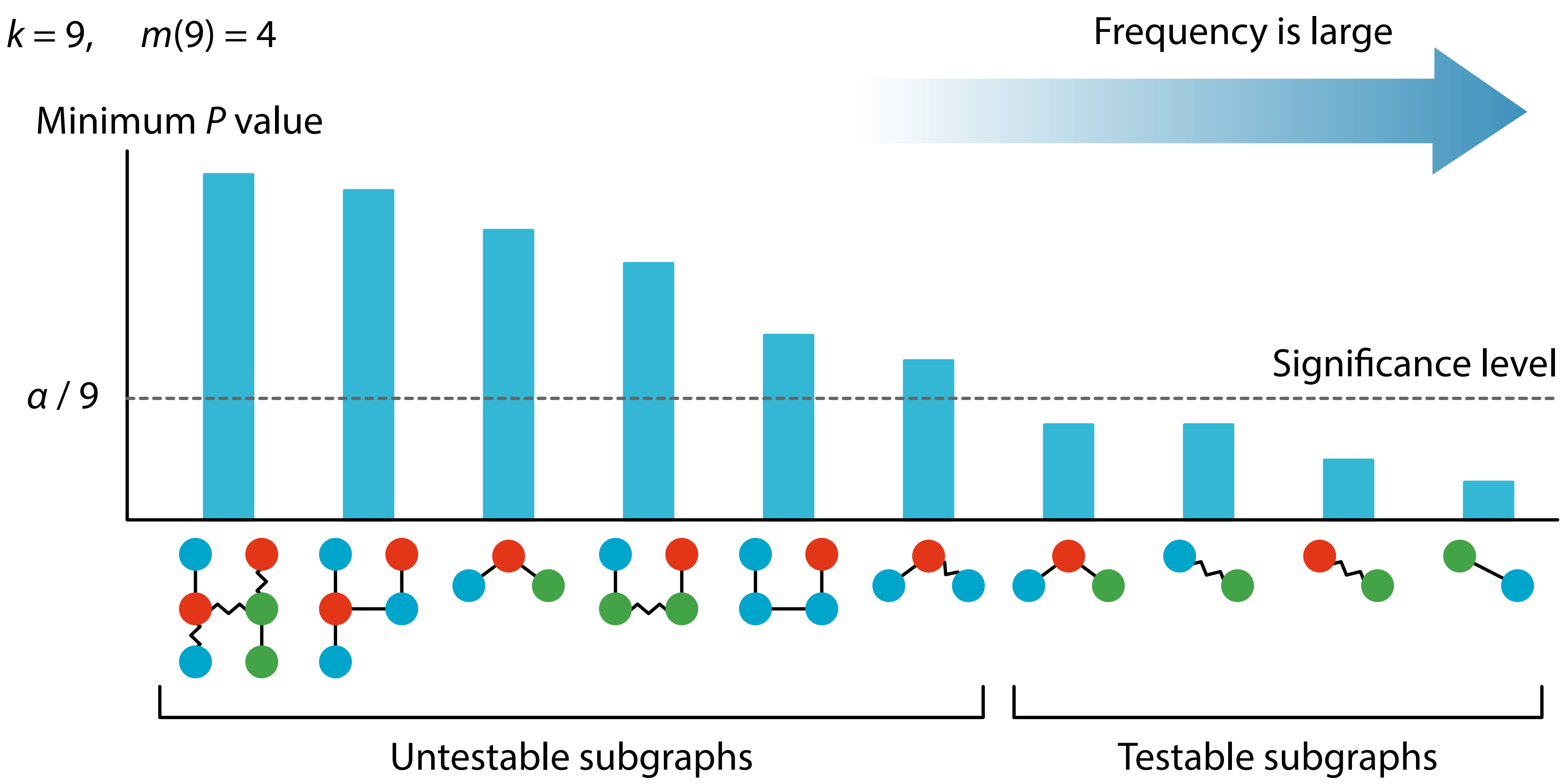
# Testable Subgraphs



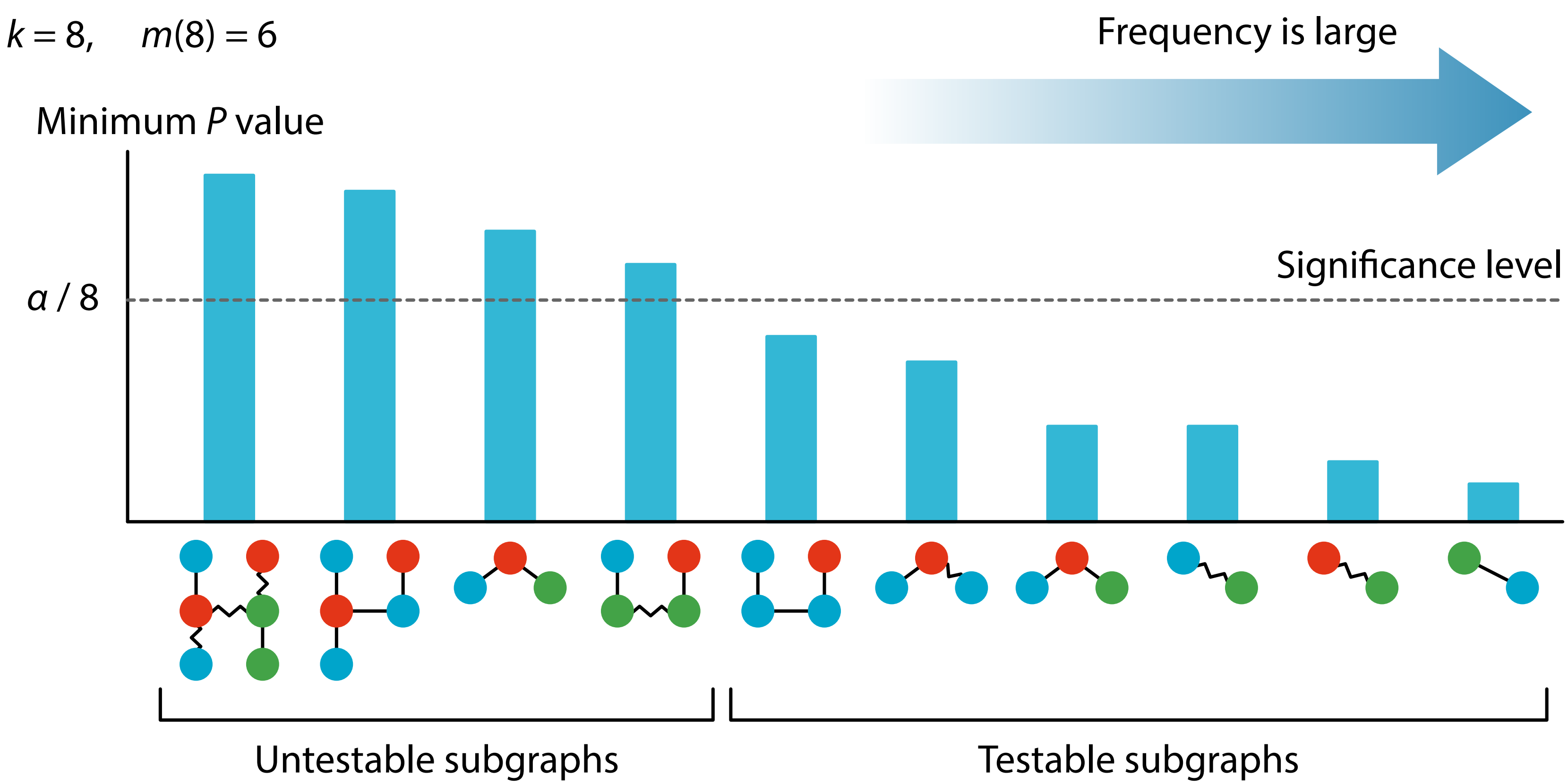
# Testable Subgraphs



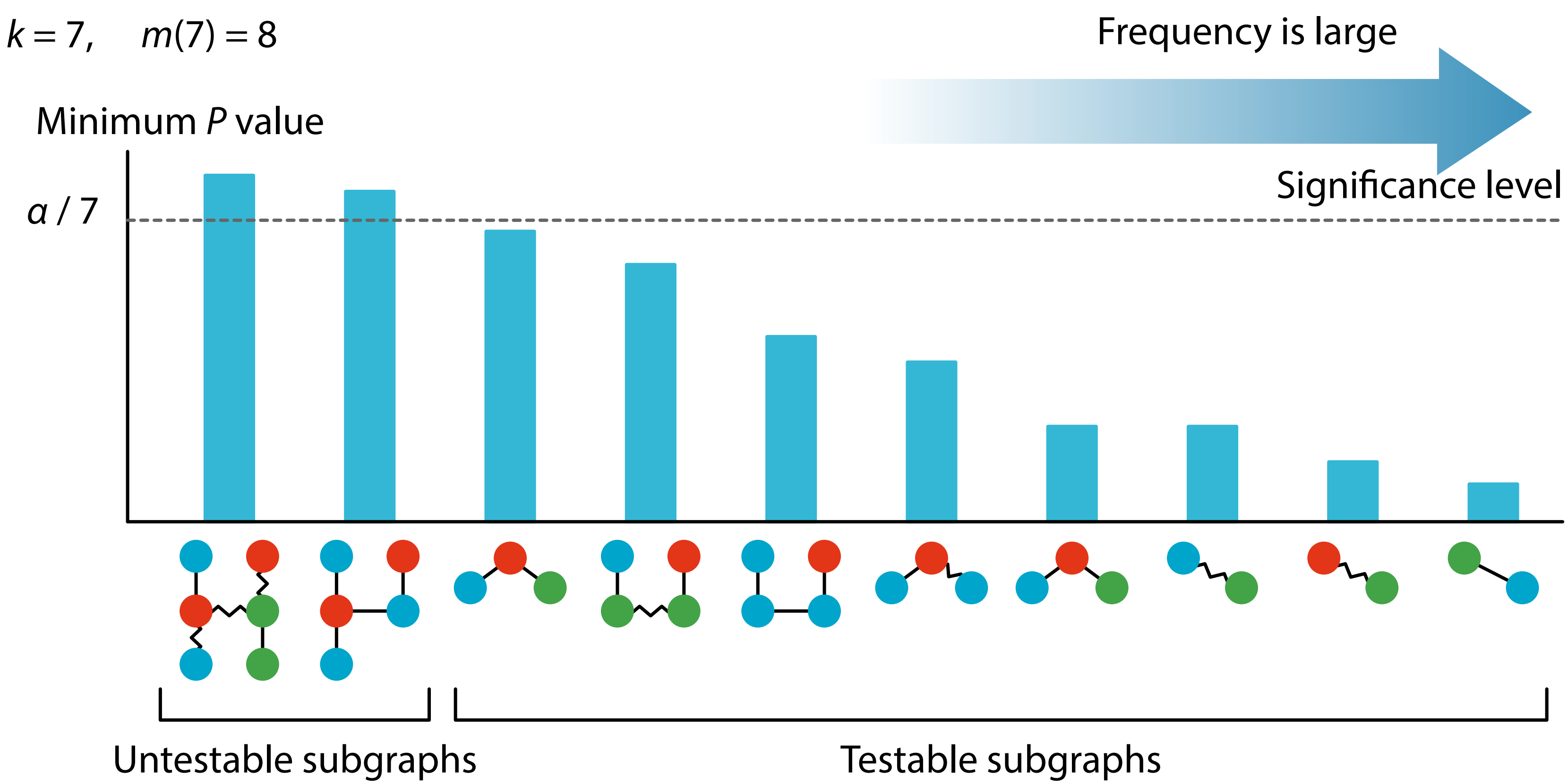
# Testable Subgraphs



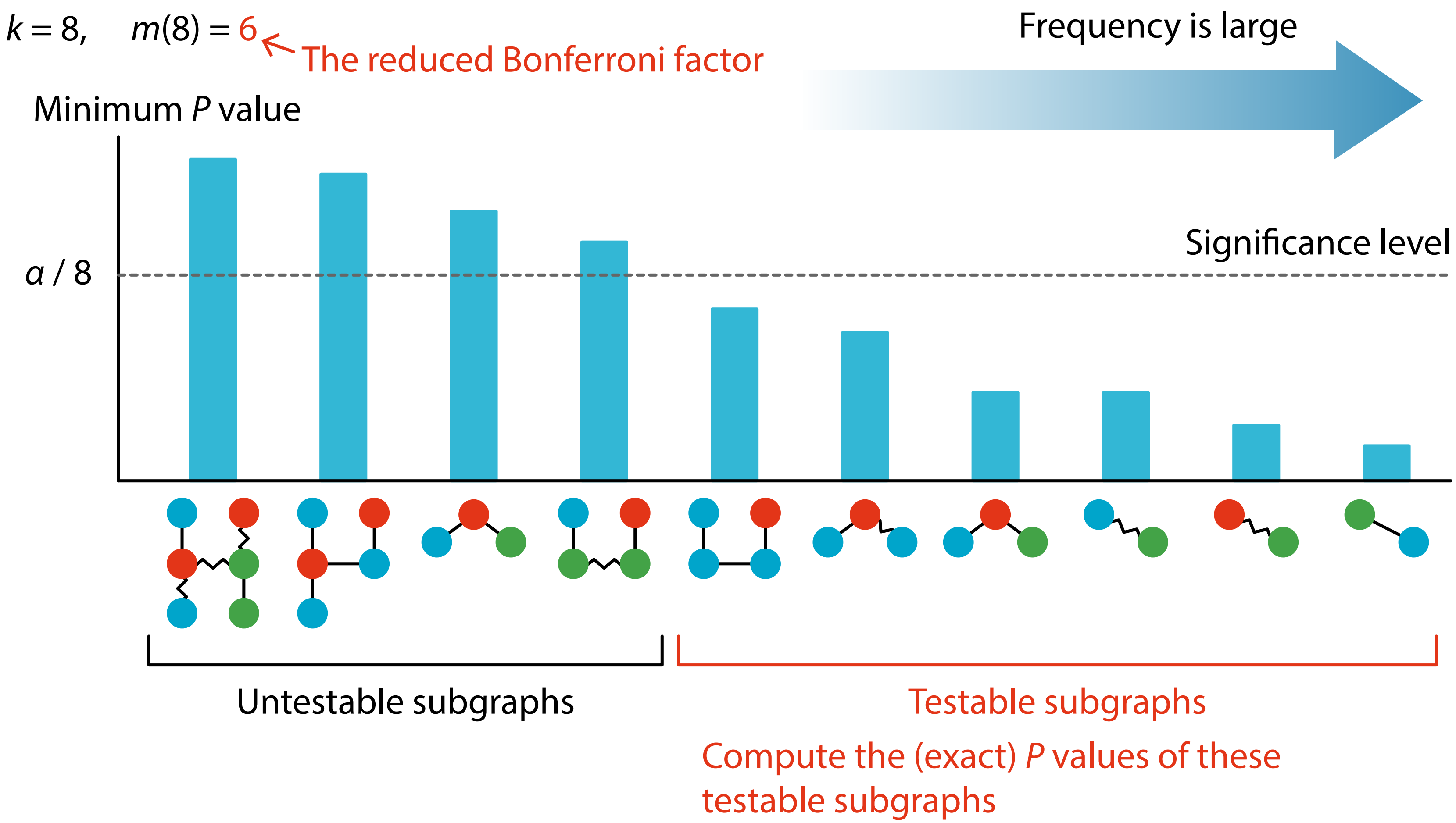
# Testable Subgraphs



# Testable Subgraphs



# Testable Subgraphs



# Effective Number of Tests

- Many subgraphs are expected to be highly correlated due to subgraph-supergraph relationships
- Use the effective number of tests to exploit the dependence between subgraphs and increase the power
- In the Šidák correction, the significance level

  $\alpha' = 1 - (1 - \alpha)^{1/m}$

  for $m$ independent tests
- Only $m_{\mathrm{eff}} < m$ tests are effective for controlling the FWER

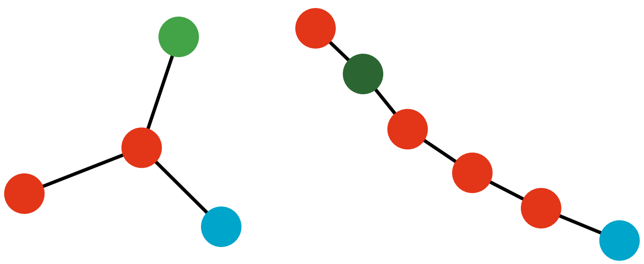
  $$m_{\mathrm{eff}} := \frac{\log(1 - \alpha)}{\log(1 - \alpha')}$$

# Estimation of Effective Number

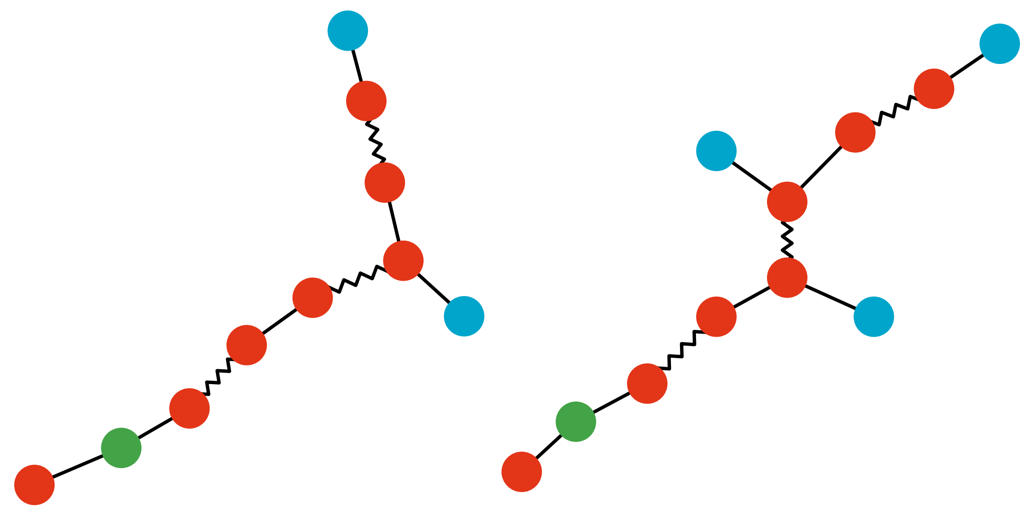
- We directly estimate the level $\alpha'$ by random permutations of class labels
  - Optimal estimation of $m_{\mathrm{eff}}$ in theory
  - The drawback is the high computational cost $O(mh)$
    - $m$: # of subgraphs, $h$: # of iterations
- Overcome by considering only testable subgraphs
  - We apply the above permutation-based estimation to only testable subgraphs
  - The complexity is $O(\tau(m)h)$ ($\tau(m)$: # of testable subgraphs)
- Moskvina, V. and Schmidt, K. M. **On multiple-testing correction in genome-wide association studies.** *Genetic epidemiology*, 32(6):567–573, 2008.

# Detected Significant Subgraphs

PTC (MR)
(carcinogenicity)



NCI 220
(anti-cancer activity)



# Frequent Subgraph Miners

- **[AGM]** Inokuchi, A. and Washio, T. and Motoda, H.: **An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data**, PKDD 2000
- **[gSpan]** Yan, X. and Han, J.: **gSpan: Graph-based substructure pattern mining**, ICDM 2002
- **[GASTON]** Nijssen, S. and Kok, J. N.: **A Quickstart in Frequent Structure Mining Can Make a Difference**, KDD 2004
- **(comparison)** Wörlein, M. and Meinl, T. and Fischer, I. and Philippsen, M. **A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, and Gaston**, PKDD 2005
  - We used GASTON as it is the fastest

# 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 $\sigma$ is too large and we decrement it