Significant Subgraph Mining with Multiple Testing Correction

Mahito Sugiyama (Osaka University, JST PRESTO)
Joint work with Felipe Llinares López¹, Niklas Kasenburg², Karsten Borgwardt¹ (¹ETH Zürich, ²Univ. 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

3/21
Find Associated Subgraphs

Active

Inactive

3/21
Find Associated Subgraphs

Active

Inactive

3/21
## 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

Fisher's exact test:
P value = 0.029

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

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

Fisher’s exact test: $P$ value = 0.029
## Multiple Testing

| Occur. | Non-occur. | Total |
|--------|------------|-------|
| Active | 1          | 3     | 4     |
| Inactive | 0        | 4     | 4     |
| Total  | 1          | 7     | 8     |

Fisher’s exact test: $P$ value = 1

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

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

Fisher’s exact test: $P$ value = 0.029
Task: Detect all significant subgraphs

Active
Inactive
Occur.
Total
Fisher's exact test: $P$ value = 0.029

Active
Inactive
Occur.
Total
Fisher's exact test: $P$ value = 0.143

Active
Inactive
Occur.
Total
Fisher's exact test: $P$ value = 1
Multiple Testing Correction

- If we test $m$ subgraphs, $am$ subgraphs are false positives
  - $\alpha$: Significance level (predetermined by the user)

- FWER: Probability of having more than one false positives among all subgraphs
  - FWER = $\Pr(FP > \alpha)$
    - 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: $\delta^{*}_{Bon} = \alpha/m$
Counting the Frequency of Subgraphs

Frequency

\[ f(\text{subgraph}) = 7 \]
Counting the Frequency of Subgraphs

Frequency
$f(\ ) = 6$

7/21
The Minimum $P$ Value

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

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

|          | 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 ($f(H) < n$)

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

$$\chi_{\text{max}} = \min\{f(H), n\}$$
Testability

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

$$P_{\text{min}} = \left( \frac{n}{f(H)} \right) \bigg/ \left( \frac{n + n'}{f(H)} \right)$$

- 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 < \( a/k \)
  - \( k \): the correction factor, \( a/k \): the corrected significance level

- For each \( k \), FWER is controlled as (Tarone 1990):
  \[
  \text{FWER} \leq m(k) \frac{a}{k} = \frac{m(k)}{k}a
  \]

- Our task is to optimize \( k \):
  \[
  k^* = \arg\max_k m(k) \text{ s.t. } m(k) \leq k
  \]
  - Enumerate testable subgraphs whose min. \( P \) values < \( a/k^* \)
    \[
    \delta_{\text{Bon}}^* = a/(\# \text{ of all subgraphs})
    \]
    \[
    \delta_{\text{Tar}}^* = a/(\# \text{ of testable subgraphs})
    \]
Subgraphs Are Testable Iff Frequent

- Our task:

\[
    k^* = \arg\max_k m(k) \quad \text{s.t. } m(k) \leq k
\]

- \( m(k) \) = \# of subgraphs whose minimum \( P \) values < \( \alpha/k \)
Subgraphs Are Testable Iff Frequent

Our task:

\[ k^* = \arg\max_k m(k) \quad \text{s.t. } m(k) \leq k \]

\[ \downarrow \]

\[ \sigma^* = \arg\max_\sigma m'(\sigma) \quad \text{s.t. } m'(\sigma) \leq \alpha / \psi(\sigma) \]

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

\[ k^* = \arg\max_k m(k) \quad \text{s.t.} \quad m(k) \leq k \]

\[ \downarrow \]

\[ \sigma^* = \arg\max_{\sigma} m'(\sigma) \quad \text{s.t.} \quad m'(\sigma) \leq \alpha / \psi(\sigma) \]

- \( m(k) \): # of subgraphs whose minimum \( P \) values < \( \alpha / k \)
- \( m'(\sigma) \): # of subgraphs whose frequency \( \geq \sigma \)
  - # 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

# of subgraphs

Frequency

12/21
Decremental Search (LAMP)

Terminate if # of subgraphs is larger than $\alpha / (P_{\text{min at } \sigma})$.
Incremental Search

Terminate if # of subgraphs detected so far exceeds $\alpha / (P_{\text{min}} \text{ at } \sigma)$
## 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

PTC(MR)

MUTAG

D&D

NCI1

NCI167

NCI220

Correction factor

10⁵

10⁷

Correction factor

10⁴

10⁶

Correction factor

10⁸

10⁹

Max. size of subgraph nodes

5 10 15 Limitless

Max. size of subgraph nodes

5 10 15 Limitless

Max. size of subgraph nodes

5 10 15 Limitless

Max. size of subgraph nodes

5 10 15 Limitless

Max. size of subgraph nodes

5 10 15 Limitless

Correction factor

Bonferroni

Testable
Number of Significant Subgraphs

- PTC(MR)
- MUTAG
- D&D
- NCI1
- NCI167
- NCI220

The graphs show the number of significant subgraphs as a function of the maximum size of subgraph nodes. The data is grouped by datasets: PTC(MR), MUTAG, D&D, NCI1, NCI167, and NCI220. Each dataset has a unique set of data points indicating the number of significant subgraphs at different node sizes, ranging from 5 to Limitless.
Running Time (second)

PTC(MR)

MUTAG

D&D

NCI1

NCI167

NCI220

Running time (s)

Max. size of subgraph nodes

5 10 15 Limitless

5 10 15 Limitless

5 10 15 Limitless

Brute-force

Decremental

Incremental
Running Time Summary

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

| Method          | 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 |  |  |  |  | Limitless |  |  |
|-----------|--------------------------------|---|---|---|---|-----------|---|---|
|           | 5 | 7 | 9 | 11 | 13 | 15 |  Limitless |  |  |
| PTC(MR)   | 9 | 10 | 11 | 11 | 11 | 11 | 11 |  |  |
| MUTAG     | 8 | 10 | 11 | 12 | 14 |  | — |  |  |
| D&D       | 20 | 22 | 22 | 22 | 22 | 22 | 22 |  |  |
| NCI1      | 17 | 20 | 22 | 25 | 27 | 29 |  | — |  |
| NCI167    | 7 | 8 | 9 | 10 | 11 |  | — |  |  |
| NCI220    | 10 | 11 | 13 | 14 | 15 | 16 | 18 |  |  |

|  | $n$ |
|---|---|
| PTC(MR) | 181 |
| MUTAG | 125 |
| D&D | 691 |
| NCI1 | 2104 |
| NCI167 | 9615 |
| NCI220 | 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 occurrence of the subgraph is **independent** from the activity

Alternative hypothesis: The occurrence 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 $H \subseteq G$ 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 formula:

$$x \cdot n' - x' \cdot n$$

Fisher's exact test table:

| Occ. | Non-occ. | Total |
|------|---------|-------|
| $x$  | $n - x$ | $n$   |
| $x'$ | $n' - x'$ | $n'$ |
| $x + x'$ | $(n - x) + (n' - x')$ | $n + n'$ |

A-4/A-16
Fisher's Exact Test

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

\[
q(x) = \frac{\binom{n}{x} \binom{n'}{x'}}{\binom{n + n'}{x + x'}}
\]

| Occ. | Non-occ. | Total |
|------|----------|-------|
| \( G \) | \( x \) | \( n - x \) | \( n \) |
| \( G' \) | \( x' \) | \( n' - x' \) | \( n' \) |
| Total | \( x + x' \) | \( n + n' \) |  |

Probability \( q(x) \)

\[
\begin{align*}
\text{max} & = \max \{0, x + x' - n'\} \\
\text{min} & = \min \{x + x', n\}
\end{align*}
\]

\( P \text{ value} \)

A-5/A-16
Testable Subgraphs

Minimum P value

Frequency is large

A-6/A-16
Testable Subgraphs

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

Minimum $P$ value

$\alpha / 10$

Frequencty is large

Significance level

Untestable subgraphs

Testable subgraphs
Testable Subgraphs

\[ k = 9, \quad m(9) = 4 \]

Minimum P value

Frequency is large

Significance level \( \alpha / 9 \)

Untestable subgraphs

Testable subgraphs
Testable Subgraphs

$k = 8, \ m(8) = 6$

Minimum $P$ value

$\alpha / 8$

Frequency is large

Significance level

Untestable subgraphs

Testable subgraphs
Testable Subgraphs

\[ k = 7, \quad m(7) = 8 \]

Minimum \( P \) value

\[ \alpha / 7 \]

Frequency is large

Significance level

Untestable subgraphs

Testable subgraphs
Testable Subgraphs

Minimum $P$ value

Significance level

$k = 8$, $m(8) = 6$

The reduced Bonferroni factor

Testable subgraphs

Compute the (exact) $P$ values of these testable subgraphs

Untestable subgraphs

Frequency is large
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
  \[ a' = 1 - (1 - a)^{1/m} \]
  for \( m \) independent tests.
- Only \( m_{\text{eff}} \) tests are effective for controlling the FWER.
  \[ m_{\text{eff}} := \frac{\log(1 - \alpha)}{\log(1 - a')} \]
Estimation of Effective Number

• We directly estimate the level $a'$ by random permutations of class labels
  – Optimal estimation of $m_{\text{eff}}$ 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 $|\mathcal{I}(\sigma)| \leq a/\psi(\sigma)$ is checked
    ○ $\mathcal{I}(\sigma)$: the set of itemsets found so far with the frequency $\geq \sigma$
  – As soon as $|\mathcal{I}(\sigma)| > a/\psi(\sigma)$, the current $\sigma$ is too large and we decrement it