Supplementary Notes for Significant Subgraph Mining with Multiple Testing Correction

Mahito Sugiyama[*][†][‡] Felipe Llinares López[§] Niklas Kasenburg[¶] Karsten M. Borgwardt[§]

A Additional Information about Datasets

The PTC (Predictive Toxicology Challenge) dataset[1] contains data of 601 chemical compounds in total (including training and test sets), which is originally designed for a prediction challenge of carcinogenic effects. Graphs are classified according to their carcinogenicity assayed on rats and mice. We assume that graphs labeled as CE, SE, or P as positive, and those of NE or N as negative, the same setting as in [4, 9]. The dataset is divided into four overlapping subsets according to their animal models: male rats (MR), female rats (FR), male mice (MM), and female mice (FM). We used only MR since the properties of other datasets are similar.

MUTAG [2] is a dataset of 188 mutagenic aromatic and heteroaromatic nitro compounds, which are classified into two classes of mutagenically active or inactive on the bacterium *Salmonella typhimurium*.

ENZYMES is a dataset of protein tertiary structures used in [1], which consists of 600 enzymes, extracted from the BRENDA database [6]. Each enzyme is classified into one of six Enzyme Commission top level enzyme classes (EC1 to EC6). We classified enzymes from EC1 to EC3 to one class, and from EC4 to EC6 to the other for our binary classification problem.

D&D is a dataset of 1178 protein structures created by Dobson and Doig [3], and they are classified into enzymes and non-enzymes. As we can see in Table 1, the size of each graph in this dataset is relatively large compared to the other datasets[2].

NCI (National Cancer Institute) datasets contain data of chemical compounds that are classified according to their anti-cancer activity [8]. Datasets are numbered by their bioassay IDs. NCI1 is balanced subsets, which is often used in the literature [5, 7], and the others are the full sets retrieved from the official website[3].

References

[1] Borgwardt, K.M., Ong, C.S., Schönauer, S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H.P.: Protein function prediction via graph kernels. Bioinformatics 21(suppl 1), i47–i56 (2005)

[2] Debnath, A.K., Lopez de Compadre, R.L., Debnath, G., Shusterman, A.J., Hansch, C.: Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. Journal of Medicinal Chemistry 34(2), 786–797 (1991)

[3] Dobson, P.D., Doig, A.J.: Distinguishing enzyme structures from non-enzymes without alignments. Journal of Molecular Biology 330(4), 771–783 (2003)

[4] Kong, X., Yu, P.S.: Semi-supervised feature selection for graph classification. In: KDD. pp. 793–802 (2010)

[5] Li, G., Semerci, M., Yener, B., Zaki, M.J.: Effective graph classification based on topological and label attributes. Statistical Analysis and Data Mining 5(4), 265–283 (2012)

[6] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D.: BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Research 32(suppl 1), D431–D433 (2004)

[7] Shervashidze, N., Schweitzer, P., van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-Lehman graph kernels. JMLR 12, 2359–2561 (2011)

[8] Wale, N., Watson, I.A., Karypis, G.: Comparison of descriptor spaces for chemical compound retrieval and classification. Knowledge and Information Systems 14(3), 347–375 (2008)

[9] Zhao, Y., Kong, X., Yu, P.S.: Positive and unlabeled learning for graph classification. In: ICDM. pp. 962–971 (2011)

[*] ISIR, Osaka University

[†] JST, PRESTO

[‡] Contact: mahito@ar.sanken.osaka-u.ac.jp

[§] D-BSSE, ETH Zürich

[¶] Department of Computer Science, University of Copenhagen

[1] http://www.predictive-toxicology.org/ptc/

[2] MUTAG, ENZYMES, and D&D are obtained from http://mlcb.is.tuebingen.mpg.de/Mitarbeiter/Nino/Graphkernels/data.zip

[3] https://pubchem.ncbi.nlm.nih.gov/

Table S1: Notation.

G, H	Graph
$V(G)$	The set of vertices of G
$E(G)$	The set of edges of G
$H \sqsubseteq G$	H is a subgraph of G
$\mathcal{G}, \mathcal{G}'$	A set of graphs
$\mathcal{H}$	The set of subgraphs in $\mathcal{G} \cup \mathcal{G}'$: $\mathcal{H} = \{H \sqsubseteq G \mid G \in \mathcal{G} \cup \mathcal{G}'\}$
$\|X\|$	Cardinality of X
n (resp. n')	Cardinality of $\mathcal{G}$ (resp. $\mathcal{G}'$): $n = \|\mathcal{G}\|$ and $n' = \|\mathcal{G}'\|$
x (resp. x')	Frequency of H in $\mathcal{G}$ (resp. $\mathcal{G}'$): $x = \|\{ G \in \mathcal{G} \mid H \sqsubseteq G \}\|$
$q(x)$	Probability $\binom{n}{x}\binom{n'}{x'}/\binom{n+n'}{x+x'}$
$f(H)$	Frequency of H in $\mathcal{G} \cup \mathcal{G}'$: $f(H) = x + x' = \|\{ G \in \mathcal{G} \cup \mathcal{G}' \mid H \sqsubseteq G \}\|$
σ	Frequency
$\psi(\sigma)$	Minimum P value of frequency σ: $\psi(\sigma) = \binom{n}{\sigma}/\binom{n+n'}{\sigma}$
$\mathcal{H}$	The set of subgraphs in $\mathcal{G} \cup \mathcal{G}'$, $\|\mathcal{H}\|$ is the Bonferroni correction factor
α	Significance level
k	Natural number
$m(k)$	The value $\|\{ H \in \mathcal{H} \mid \psi \circ f(H) \leq \alpha/k \}\|$
k_{rt}	(Rounded) Root of $m(k) - k$: $m(k_{\mathrm{rt}} - 1) > k_{\mathrm{rt}} - 1$, $m(k_{\mathrm{rt}}) \leq k_{\mathrm{rt}}$
$\tau(\mathcal{H})$	The set of testable subgraphs: $\tau(\mathcal{H}) = \{ H \in \mathcal{H} \mid \psi \circ f(H) \leq \alpha/k_{\mathrm{rt}} \}$
σ_{rt}	(Rounded) Root frequency such that $\|\{ H \in \mathcal{H} \mid f(H) \geq (\sigma_{\mathrm{rt}} - 1) \}\| > \alpha/\psi(\sigma_{\mathrm{rt}} - 1)$ and $\|\{ H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}} \}\| \leq \alpha/\psi(\sigma_{\mathrm{rt}})$
$\sigma_{\min}$	The minimum possible frequency $\sigma_{\min}$ satisfying $\psi(\sigma_{\min}) < \alpha$
$\sigma_{\max}$	The maximum possible frequency n
$s(\mathcal{H})$	The set of significant subgraphs
m_{eff}	The effective number of tests within the testable subgraphs

Table S2: Root frequencies σ_{rt} for each dataset and each maximum size of subgraph nodes. "—" means that computation did not finished and the root frequency is not confirmed.

	Maximum size of subgraph nodes													
Dataset	4	5	6	7	8	9	10	11	12	13	14	15	16	Limitless
PTC(MR)	9	9	10	10	11	11	11	11	11	11	11	11	11	11
MUTAG	8	8	9	10	10	11	12	12	13	14	—	—	—	—
ENZYMES	11	14	15	17	19	22	24	27	—	—	—	—	—	—
D&D	17	20	21	22	22	22	22	22	22	22	22	22	22	22
NCI1	16	17	19	20	21	22	24	25	26	27	28	29	30	—
NCI41	5	5	6	6	6	7	7	7	8	8	8	—	—	—
NCI167	6	7	7	8	8	9	9	10	10	11	—	—	—	—
NCI220	9	10	11	11	12	13	13	14	14	15	15	16	16	18

Algorithm 1 One-pass search

```
Input: Datasets $\mathcal{G}$, $\mathcal{G}'$ and significance level $\alpha$
Output: All significant subgraphs
$\sigma_{\min} \leftarrow 1$
while $\psi(\sigma_{\min}) > \alpha$ do
  $\sigma_{\min} \leftarrow \sigma_{\min} + 1$
end while
// $\sigma_{\min}$ is the minimum possible frequency
$\mathcal{H}(\sigma_{\min}) \leftarrow \{H \in \mathcal{H} \mid f(H) \geq \sigma_{\min}\}$
// This set is obtained by running an FSM
// algorithm with the threshold $\sigma_{\min}$
$\sigma_{\mathrm{rt}} \leftarrow \sigma_{\min}$
while $|\,\{\, H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}} \,\}\,| > \alpha/\psi(\sigma_{\mathrm{rt}})$ do
  $\sigma_{\mathrm{rt}} \leftarrow \sigma_{\mathrm{rt}} + 1$
end while
// $\sigma_{\mathrm{rt}}$ is the root frequency
$\tau(\mathcal{H}) \leftarrow \{\, H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}} \,\}$
// Testable hypotheses
$s(\mathcal{H}) \leftarrow \{\, H \in \tau(\mathcal{H}) \mid P$ value of $H \;< \alpha/|\tau(\mathcal{H})| \,\}$
Output $s(\mathcal{H})$
```

Algorithm 3 Incremental search

```
Input: Datasets $\mathcal{G}$, $\mathcal{G}'$ and significance level $\alpha$
Output: All significant subgraphs
$\sigma_{\mathrm{rt}} \leftarrow 1$
while $\psi(\sigma_{\mathrm{rt}}) > \alpha$ do
  $\sigma_{\mathrm{rt}} \leftarrow \sigma_{\mathrm{rt}} + 1$
end while
// This is the minimum possible frequency
repeat
  Run an FSM algorithm with the threshold $\sigma_{\mathrm{rt}}$
  with monitoring the number $m$ of frequent subgraphs
  if $m > \alpha/\psi(\sigma_{\mathrm{rt}})$ while the process then
    Terminate the mining process
  else
    $\mathcal{H}(\sigma_{\mathrm{rt}}) \leftarrow \{H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}}\}$
    // This set is obtained by running an FSM
    // algorithm with the threshold $\sigma_{\mathrm{rt}}$
  end if
  $\sigma_{\mathrm{rt}} \leftarrow \sigma_{\mathrm{rt}} + 1$
until the mining process is not terminated
$\sigma_{\mathrm{rt}} \leftarrow \sigma_{\mathrm{rt}} - 1$    // $\sigma_{\mathrm{rt}}$ is the root frequency
$\tau(\mathcal{H}) \leftarrow \{\, H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}} \,\}$
// Testable hypotheses
$s(\mathcal{H}) \leftarrow \{\, H \in \tau(\mathcal{H}) \mid P$ value of $H \;< \alpha/|\tau(\mathcal{H})| \,\}$
Output $s(\mathcal{H})$
```

Algorithm 2 Decremental search (LAMP search)

```
Input: Datasets $\mathcal{G}$, $\mathcal{G}'$ and significance level $\alpha$
Output: All significant subgraphs
$\sigma_{\mathrm{rt}} \leftarrow n$    // the maximum possible frequency
repeat
  $\mathcal{H}(\sigma_{\mathrm{rt}}) \leftarrow \{H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}}\}$
  // This set is obtained by running an FSM
  // algorithm with the threshold $\sigma_{\mathrm{rt}}$
  $\sigma_{\mathrm{rt}} \leftarrow \sigma_{\mathrm{rt}} - 1$
until $|\mathcal{H}(\sigma_{\mathrm{rt}})| > \alpha/\psi(\sigma_{\mathrm{rt}})$
$\sigma_{\mathrm{rt}} \leftarrow \sigma_{\mathrm{rt}} + 2$    // $\sigma_{\mathrm{rt}}$ is the root frequency
$\tau(\mathcal{H}) \leftarrow \{\, H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}} \,\}$
// Testable hypotheses
$s(\mathcal{H}) \leftarrow \{\, H \in \tau(\mathcal{H}) \mid P$ value of $H \;< \alpha/|\tau(\mathcal{H})| \,\}$
Output $s(\mathcal{H})$
```

Algorithm 4 Bisection search (LEAP search)

```
Input: Datasets $\mathcal{G}$, $\mathcal{G}'$ and significance level $\alpha$
Output: All significant subgraphs
$\sigma_{\min} \leftarrow 1$
while $\psi(\sigma_{\min}) > \alpha$ do
  $\sigma_{\min} \leftarrow \sigma_{\min} + 1$
end while
$\sigma_{\max} \leftarrow n$    // the maximum possible frequency
$\sigma_{\mathrm{rt}} \leftarrow \lfloor(\sigma_{\min} + \sigma_{\max})/2\rfloor$
repeat
  Run an FSM algorithm with the threshold $\sigma_{\mathrm{rt}}$
  with monitoring the number $m$ of frequent subgraphs
  if $m > \alpha/\psi(\sigma_{\mathrm{rt}})$ while the process then
    Terminate the mining process
  else
    $\mathcal{H}(\sigma_{\mathrm{rt}}) \leftarrow \{H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}}\}$
  end if
  if the mining process is terminated then
    $\sigma_{\min} \leftarrow \sigma_{\mathrm{rt}}$
  else
    $\sigma_{\max} \leftarrow \sigma_{\mathrm{rt}}$
  end if
  $\sigma_{\mathrm{rt}} \leftarrow \lfloor(\sigma_{\min} + \sigma_{\max})/2\rfloor$
until $\sigma_{\max} - \sigma_{\min} = 1$
if the last mining process was terminated then
  $\sigma_{\mathrm{rt}} \leftarrow \sigma_{\max}$    // $\sigma_{\mathrm{rt}}$ is the root frequency
end if
$\tau(\mathcal{H}) \leftarrow \{\, H \in \mathcal{H} \mid f(H) \geq \sigma_{\mathrm{rt}} \,\}$
// Testable hypotheses
$s(\mathcal{H}) \leftarrow \{\, H \in \tau(\mathcal{H}) \mid P$ value of $H \;< \alpha/|\tau(\mathcal{H})| \,\}$
Output $s(\mathcal{H})$
```

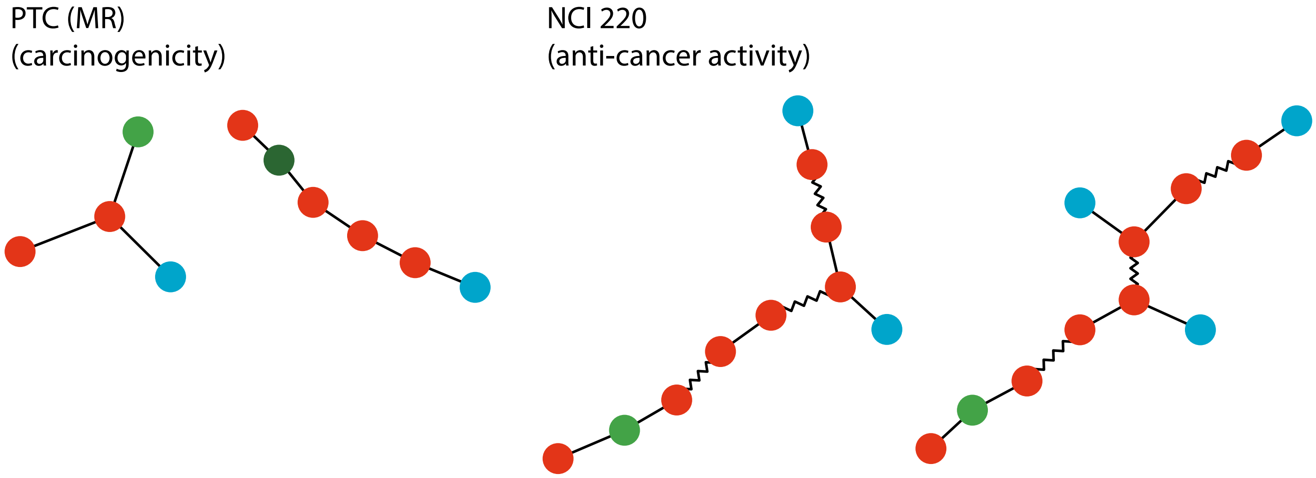

Figure S1: Four examples of significant subgraphs on PTC(MR) (left) and NCI220 (right) that are detected by our method using the testability but are missed by the standard Bonferroni factor. Different colors (resp. shapes) of vertices (resp. edges) mean different labels of them.