# Improved Algorithms for Enumerating Tree-like Chemical Graphs with Given Path Frequency

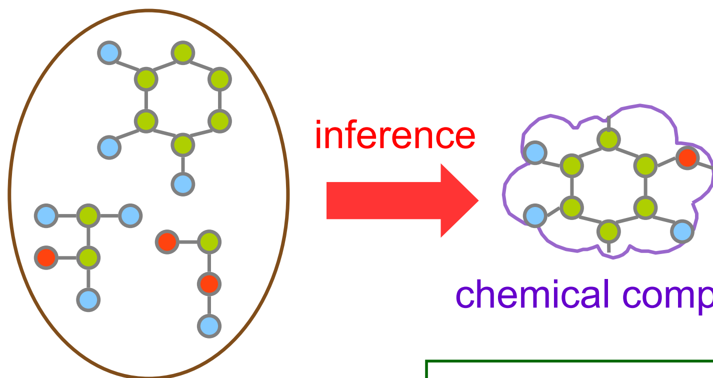
Yusuke Ishida* Liang Zhao Hiroshi Nagamochi
Graduate School of Informatics, Kyoto University, Japan

Tatsuya Akutsu
Institute for Chemical Research, Kyoto University, Japan

## Outline

1. Problem Formulation
2. Canonical Representation and Family tree
3. Algorithm (Branch and Bound)
   – Detachment-cut
4. Experimental Results for the first formulation
5. H-less Single-bond Formulation
   – Hydrogen-cut
6. Experimental Results for the second formulation
7. Conclusions

## Background



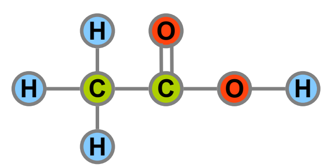

given partial structures

applications:

- structure determination using mass-spectrum
- drug design

## Definition

A feature vector $f_K(G)$ :
#occurrences of each vertex-labeled path of length 0,1,..., $K$




a chemical compound $G$

$f_3(G)$ ($K$ = 3)

length = 0

| | |
|---|---|
| H | 4 |
| O | 2 |
| C | 2 |

length = 1

| | |
|---|---|
| OH | 1 |
| CH | 3 |
| CO | 2 |
| CC | 1 |

length = 2

| | |
|---|---|
| COH | 1 |
| HCH | 3 |
| CCH | 3 |
| OCO | 1 |
| CCO | 2 |

length = 3

| | |
|---|---|
| OCOH | 1 |
| CCOH | 1 |
| OCCH | 6 |

# Enumerating Chemical Multitree Problem

**Input**

label set
$\Sigma$ = { H, O, C }

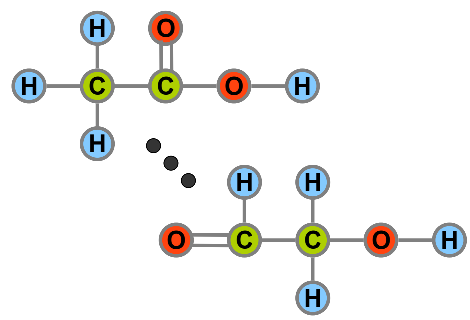
valence
val (H) = 1
val (O) = 2
val (C) = 4

level of *g*
$K = 1$

feature vector *g*

| | |
|---|---|
| H | 4 |
| O | 2 |
| C | 2 |
| OH | 1 |
| CH | 3 |
| CO | 2 |
| CC | 1 |

**Output**

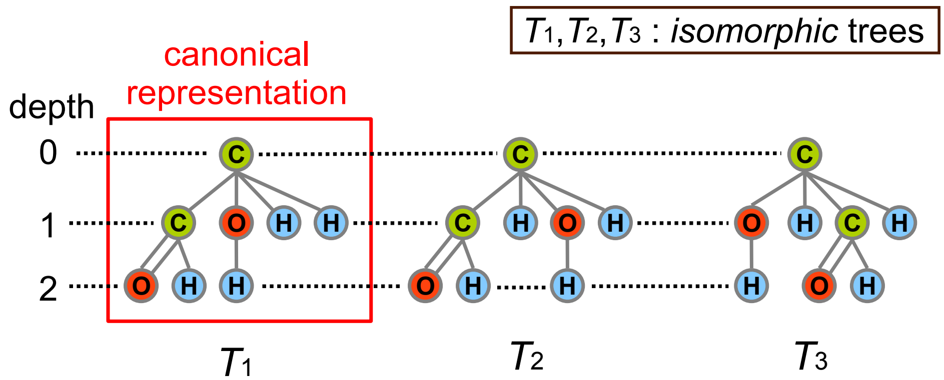



all $\Sigma$-labeled multitrees satisfying the feature vector constraint and the valence constraint

# Previous Work

- Aringhieri et al. [4OR, 2003]
  – designed two algorithms to generate all alkane isomers.
- Fujiwara et al. [J. Chem. Inf. Model., 2008]
  – proposed a branch and bound algorithm for chemical multitree problem.
  – gave H-less single-bond formulation of chemical multitree problem and their algorithm can be also applied to it.

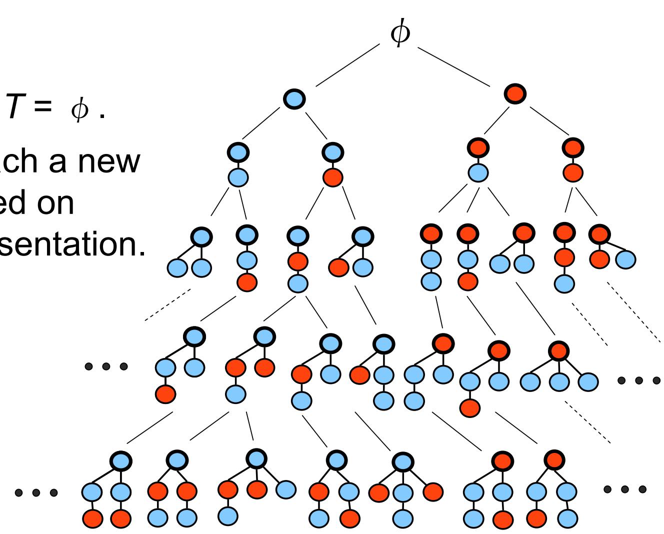
## Canonical Representation for multitrees




Define the canonical representation as
the embedding with largest depth-label sequence.

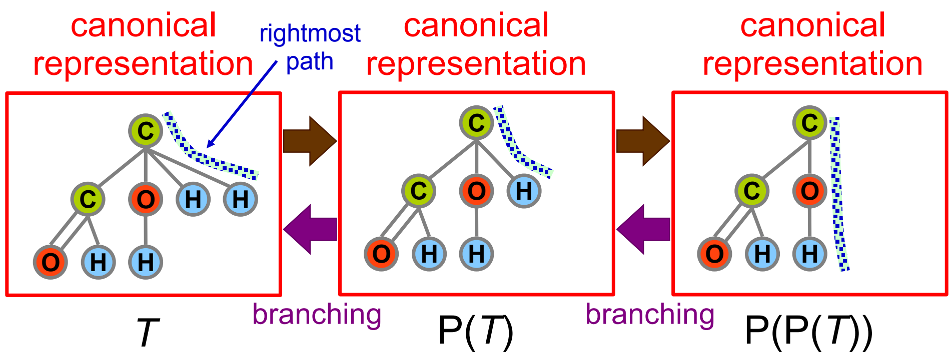
## Family tree

- Add a vertex to $T = \phi$ .
- Repeatedly attach a new vertex to $T$ based on canonical representation.




family tree（#labels=2, #vertices=5 ）

## Parent of multitrees




Define the parent P(*T*) of *T* as the tree obtained by removing the rightmost leaf from *T*.

**If *T* is canonical, P(*T*) remains canonical.**

Branching operation : *O*(1) time / node

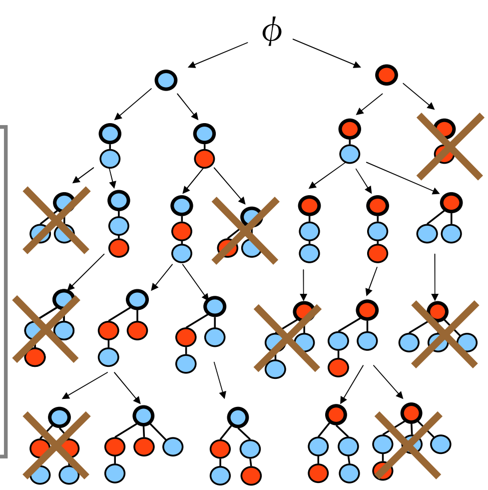
## Algorithm (Branch and Bound)

Add a vertex to $T = \phi$ .

Attach a new vertex to *T* based on canonical representation. (Branching Operation)

Check each constraint and decide whether to discard *T*. (Bounding Operation)

If $|T|$ = (#input vertices), output *T* as a solution.




family tree（#labels=2, #vertices=5）

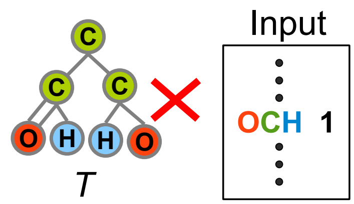
## Bounding Operations

For a current tree $T$,

①feature-vector-cut

test whether $f_K(T) \leqq g$ holds.

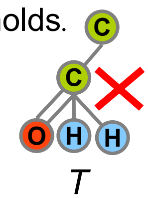
the feature vector of $T$ the input




②bond-cut

test whether $deg(v;T) \leqq val(\ell(v))$ for $\forall v \in T$ holds.

#edges incident to $v$ in $T$ the valence of the label of $v$




③detachment-cut

test whether $T$ can be extended to multitrees satisfying degree constraint.

## Detachment-cut

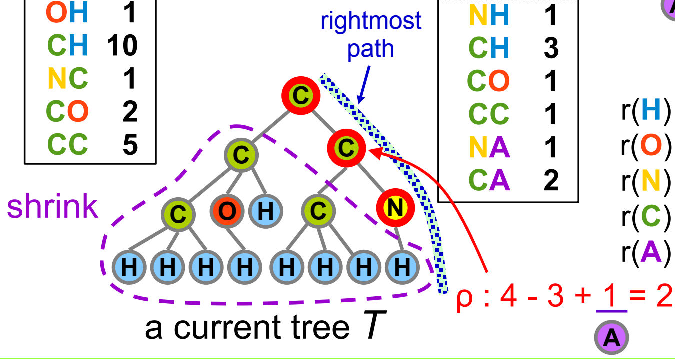
Test whether $T$ can be extended to multitrees satisfying degree constraint.

input

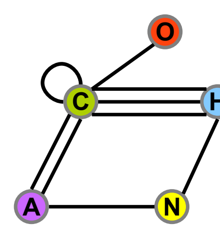
| | |
|---|---|
| H | 13 |
| O | 2 |
| N | 1 |
| C | 6 |
| NH | 2 |
| OH | 1 |
| CH | 10 |
| NC | 1 |
| CO | 2 |
| CC | 5 |

residual

| | |
|---|---|
| H | 4 |
| O | 1 |
| N | 1 |
| C | 3 |
| A | 1 |
| NH | 1 |
| CH | 3 |
| CO | 1 |
| CC | 1 |
| NA | 1 |
| CA | 2 |




a current tree $T$

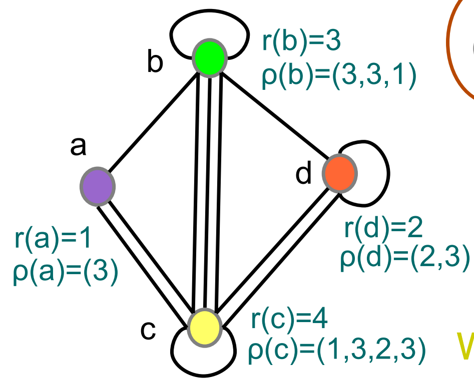


$G=(V, E)$

r(H) = 4 ρ(H) = (1,1,1,1)
r(O) = 1 ρ(O) = (2)
r(N) = 1 ρ(N) = (2)
r(C) = 3 ρ(C) = (4,3,2)
r(A) = 1 ρ(A) = (3)

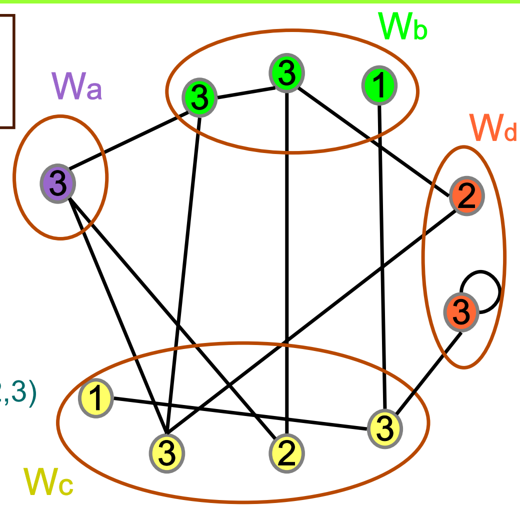
## Detachment : A Reverse Operation of Contraction

The number r($v$) of copies of $v$ and the degree ρ($v^i$) of each copy $v^i$ are specified.




a multigraph $G$

$G = (V, E)$




a ρ-detachment $H$ of $G$

$H = (\cup_{v \in V} W_v, E)$

## Detachment-cut

Test whether G has a connected and loopless ρ-detachment.

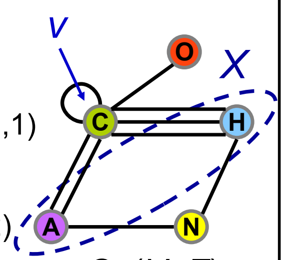
4+3+2 = 9    8    for $v$ = C

① $\sum_{1 \leqq i \leqq r(v)} \rho(v^i) \geqq deg(v;G)$   $\forall v \in V$

#edges incident to $v$ in $G$

condition for degree constraint

ρ(H) = (1,1,1,1)
ρ(O) = (2)
ρ(N) = (2)
ρ(C) = (4,3,2)
ρ(A) = (3)




$G$=($V$, $E$)

7    5    2

② $d(X,V;G) \geqq \sum_{v \in X} r(v) + c(G\text{-}X) - 1$   $\phi \neq \forall X \subseteq V$

#edges joining $X$ and $V$    #connected components in $G$ - $X$

condition for connectivity

r(H) = 4
r(O) = 1
r(N) = 1
r(C) = 3
r(A) = 1

## Experiment

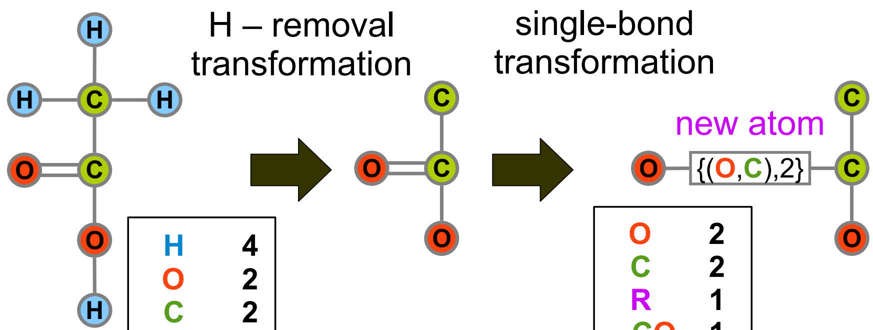
- Comparing the running time of our algorithm with Fujiwara et al.'s [2008]
- Instances from KEGG LIGAND database
  (Replacing each benzene ring by a virtual atom of valence 6)
- Pentium4 3.00GHz
- T.O. : time over 1800 (sec)

## Experimental Results for the first formulation

N.F. : not found

| Formula #atoms | *K* | Fujiwara et al.'s algorithm CPU time (sec) | Fujiwara et al.'s algorithm #solutions | our algorithm CPU time (sec) | our algorithm #solutions |
|---|---|---|---|---|---|
| $C_{16}H_{22}O_4$ | 1 | T.O. | N.F. | 158.23 | 570,773 |
| 37 | 2 | 3.11 | 9 | 0.48 | 9 |
| | 3 | 3.25 | 2 | 0.30 | 2 |
| $C_{17}H_{28}N_2O$ | 1 | T.O. | N.F. | 109.27 | 73,711 |
| 43 | 2 | 50.55 | 55 | 1.40 | 55 |
| | 3 | 16.78 | 1 | 0.61 | 1 |
| $C_{21}H_{28}N_2O_5$ | 1 | T.O. | N.F. | 500.78 | 70,170 |
| 46 | 2 | 51.72 | 16 | 3.51 | 16 |
| | 3 | 4.26 | 2 | 0.32 | 2 |
| $C_{24}H_{38}O_4$ | 1 | T.O. | N.F. | T.O. | N.F. |
| 61 | 2 | T.O. | N.F. | 318.68 | 1,198 |
| | 3 | T.O. | N.F. | 188.13 | 8 |

## H-less Single-bond Formulation [Fujiwara et al. 2008]




| | |
|---|---|
| H | 4 |
| O | 2 |
| C | 2 |
| OH | 1 |
| CH | 3 |
| CO | 2 |
| CC | 1 |

| | |
|---|---|
| O | 2 |
| C | 2 |
| R | 1 |
| CO | 1 |
| RO | 1 |
| CC | 1 |
| RC | 1 |

R = {(O,C),2}

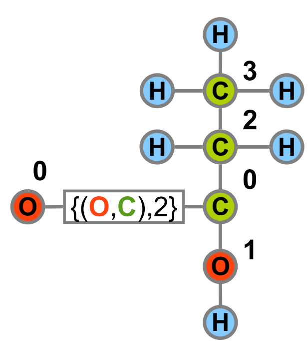
## Hydrogen-cut

*h**(ℓ) : sum of #hydrogens adjacent
to each atom of label ℓ.

input

| | |
|---|---|
| O | 2 |
| C | 3 |
| R | 1 |
| CO | 1 |
| RO | 1 |
| CC | 2 |
| RC | 1 |

R = {(O,C),2}



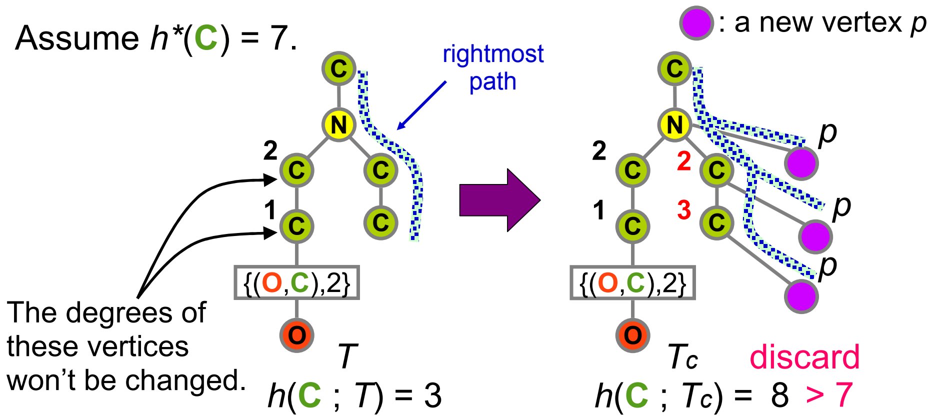

*h**(C) = 3 + 2 + 0 = 5
*h**(O) = 0 + 1 = 1

## Hydrogen-cut

$h(\ell, T)$ : sum of #hydrogens that *must* be adjacent to each atom of label $\ell$ in a current tree $T$

If $h(\ell ;T) > h^*(\ell)$ holds for a label $\ell$, discard $T$.

Assume $h^*(C) = 7$.




## Experimental Results for the second formulation

| Formula #atoms | K | Fujiwara et al.'s algorithm CPU time (sec) | Fujiwara et al.'s algorithm #solutions | our algorithm CPU time (sec) | our algorithm #solutions |
|---|---|---|---|---|---|
| $C_{21}H_{28}N_2O_5$ | 1 | 222.29 | 70,170 | 9.03 | 70,170 |
| 19 | 2 | 0.11 | 16 | 0.02 | 16 |
| | 3 | 0.09 | 2 | 0.01 | 2 |
| $C_{24}H_{38}O_4$ | 1 | T.O. | 5,305,243 | T.O. | 60,257,365 |
| 25 | 2 | 23.36 | 1,198 | 8.10 | 1,198 |
| | 3 | 15.87 | 8 | 5.66 | 8 |
| $C_{19}H_{39}O_7P$ | 2 | T.O. | 161 | 1543.37 | 2,520 |
| 29 | 3 | 184.54 | 1 | 45.36 | 1 |
| | 4 | 11.86 | 1 | 3.60 | 1 |
| $C_{21}H_{39}O_7P$ | 2 | T.O. | 77 | T.O. | 1,736 |
| 33 | 3 | T.O. | 11 | 438.19 | 13 |
| | 4 | 118.48 | 11 | 25.65 | 11 |

## Conclusions

- We proposed a branch and bound algorithm with two new bounding operations.
- For the first formulation, we can solve the problem with about 25 non-hydrogen atoms for K $\geqq$ 2.
- For the second formulation, we can solve the problem with about 30 non-hydrogen atoms for K $\geqq$ 2.

## Future Work

- Treat more general graphs (e.g., outerplanar graphs).
- Use other graph structures for representing partial structures of input.

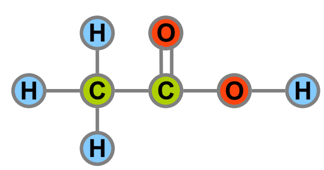
## Definition

A feature vector $f_K(G)$ :

#occurrences of each vertex-labeled path of length 0,1,..., $K$

$f_1(G)$

| | |
|---|---|
| H | 4 |
| O | 2 |
| C | 2 |
| HH | 0 |
| OH | 1 |
| CH | 3 |
| HO | 1 |
| OO | 0 |
| CO | 2 |
| HC | 3 |
| OC | 2 |
| CC | 2 |

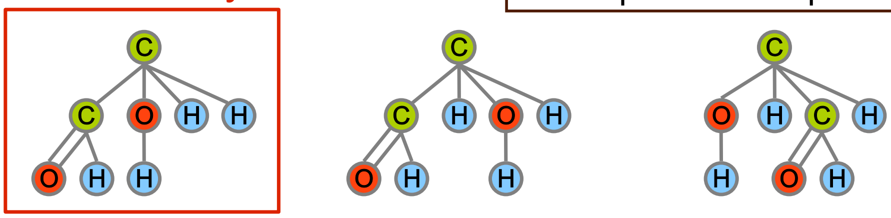



a chemical compound $G$

## Canonical Representation for multitrees

$T_1, T_2, T_3$ : *isomorphic* trees
DL: depth label sequence

left-heavy



$T_1$ DL($T_1$) = (0,C,1,C,2,O,2,H,1,O,2,H,1,H,1,H)

$T_2$ DL($T_2$) = (0,C,1,C,2,O,2,H,1,H,1,O,2,H,1,H)

$T_3$ DL($T_3$) = (0,C,1,O,2,H,1,H,1,C,2,O,2,H,1,H)

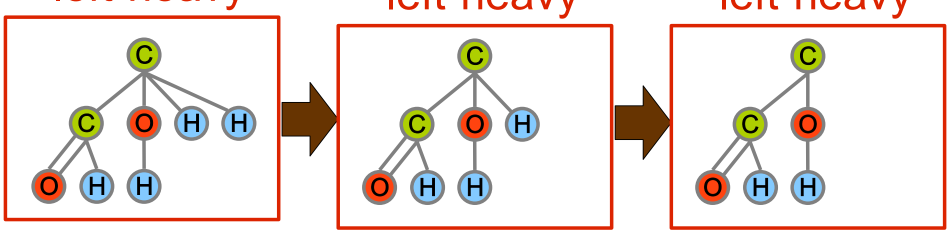
The embedding with largest DL is *left-heavy.*

[Nakano, Uno (2003)]

→ We enumerate left-heavy trees.

## Parent of multitrees

left-heavy left-heavy left-heavy



T P(T) P(P(T))

parent
P(T) : tree obtained by removing the rightmost leaf from T

If T is left-heavy, P(T) remains left-heavy.

→ In branching operation,
we attach a new vertex to rightmost path.

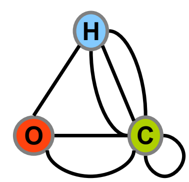
## Feature Vector ($K$=1) and ρ-detachment

Σ = { H, O, C }
val (H) = 1
val (O) = 2
val (C) = 4

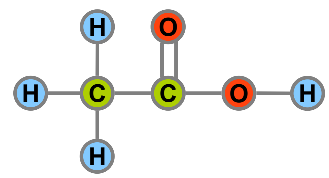
| | |
|---|---|
| H | 4 |
| O | 2 |
| C | 2 |
| HH | 0 |
| OH | 1 |
| CH | 3 |
| HO | 1 |
| OO | 0 |
| CO | 2 |
| HC | 3 |
| OC | 2 |
| CC | 2 |

r(H) = 4
r(O) = 2
r(C) = 2

ρ(H) = (1,1,1,1)
ρ(O) = (2,2)
ρ(C) = (4,4)




*G*




## Detachment-cut

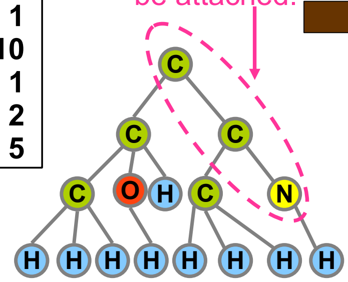
Test whether $T$ will generate trees satisfying given constraint.

input

| | |
|---|---|
| H | 13 |
| O | 2 |
| N | 1 |
| C | 6 |
| NH | 2 |
| OH | 1 |
| CH | 10 |
| NC | 1 |
| CO | 2 |
| CC | 5 |

Which vertices are counted as "residual"?

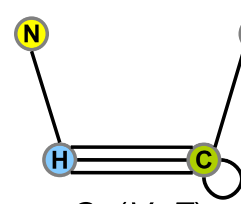
New vertices can be attached.




a current tree $T$

residual

| | |
|---|---|
| H | 4+ |
| O | 1+ |
| N | 0+ |
| C | 1+ |
| NH | 1 |
| CH | 3 |
| CO | 1 |
| CC | 1 |




$G$=($V$, $E$)

r(H)= 4+
r(O)= 1+
r(N)= 0+
r(C)= 1+

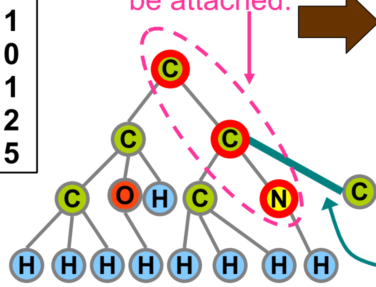
# Detachment-cut

Test whether $T$ will generate trees satisfying given constraint.

input

| | |
|---|---|
| H | 13 |
| O | 2 |
| N | 1 |
| C | 6 |
| NH | 2 |
| OH | 1 |
| CH | 10 |
| NC | 1 |
| CO | 2 |
| CC | 5 |

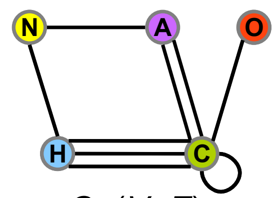
Which vertices are counted as “residual”?



a current tree $T$

residual

| | |
|---|---|
| H | 4 |
| O | 1 |
| N | 1 |
| C | 3 |
| A | 1 |
| NH | 1 |
| CH | 3 |
| CO | 1 |
| CC | 1 |
| NA | 1 |
| CA | 2 |



$G=(V, E)$

r(H)= 4
r(O)= 1
r(N)= 1
r(C)= 3
r(A)= 1

# Detachment-cut

Test whether ① and ② holds.

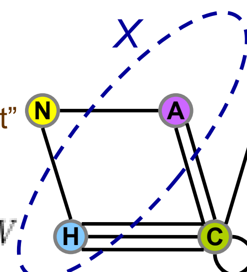
necessary conditions for
“$G$ has a connected and loopless ρ-detachment”
[Nagamochi 2006]

① $\underbrace{\overset{7}{d(X,V;G)}}_{\text{#edges joining } X \text{ and } V} \geq \overset{5}{\Sigma_{v\in X} r(v)} + \underbrace{\overset{2}{c(G-X)}}_{\text{#connected components in } G - X} - 1 \quad \emptyset \neq \forall X \subseteq V$

condition for connectivity

② $\overset{4+3+2\,=\,9}{\Sigma_{1\leq i\leq r(v)} \rho_i^v} \geq \underbrace{\overset{8}{deg(v;G)}}_{\text{#edges incident to } v \text{ in } G} \quad \forall v \in V$ for $v$ = C

condition for degree specification




$G=(V, E)$

ρ(H) = (1,1,1,1) r(H)= 4
ρ(O) = (2) r(O)= 1
ρ(N) = (2) r(N)= 1
ρ(C) = (4,3,2) r(C)= 3
ρ(A) = (3) r(A)= 1

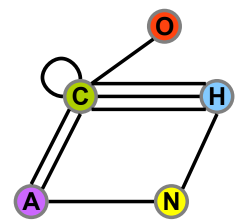
# Detachment-cut

Consider the degree constraint $\rho(v^i)$ of each vertex $v^i$ for $1 \leqq i \leqq r(v)$.

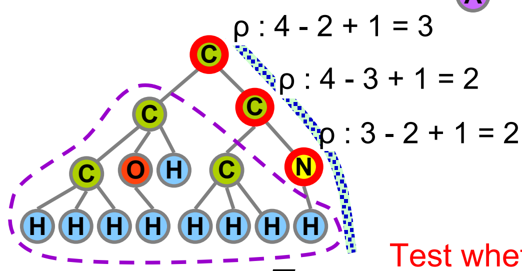
For a vertex $v^i \notin T$, $\rho(v^i) = val(\ell(v^i))$.

For a vertex $v^i \in T$,
$\rho(v^i) = val(\ell(v^i)) - deg(v^i;T) + \underline{1}$.

A




$G=(V, E)$




a current tree $T$

$\rho(\mathbf{H}) = (1,1,1,1)$  $r(\mathbf{H}) = 4$
$\rho(\mathbf{O}) = (2)$  $r(\mathbf{O}) = 1$
$\rho(\mathbf{N}) = (2)$  $r(\mathbf{N}) = 1$
$\rho(\mathbf{C}) = (4,3,2)$  $r(\mathbf{C}) = 3$
$\rho(\mathbf{A}) = (3)$  $r(\mathbf{A}) = 1$

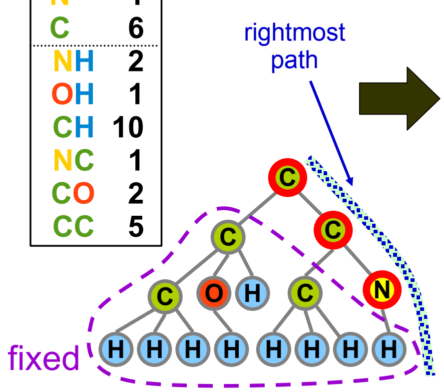
Test whether there exists multitrees satisfying the degree constraint by considering a "*detachment*" of *G*.

# Detachment-cut

Test whether *T* can be extended multitrees satisfying degree constraint.

input

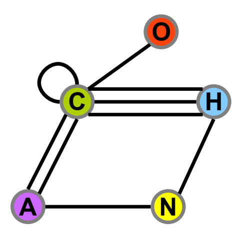
| | |
|---|---|
| H | 13 |
| O | 2 |
| N | 1 |
| C | 6 |
| NH | 2 |
| OH | 1 |
| CH | 10 |
| NC | 1 |
| CO | 2 |
| CC | 5 |




a current tree $T$

residual

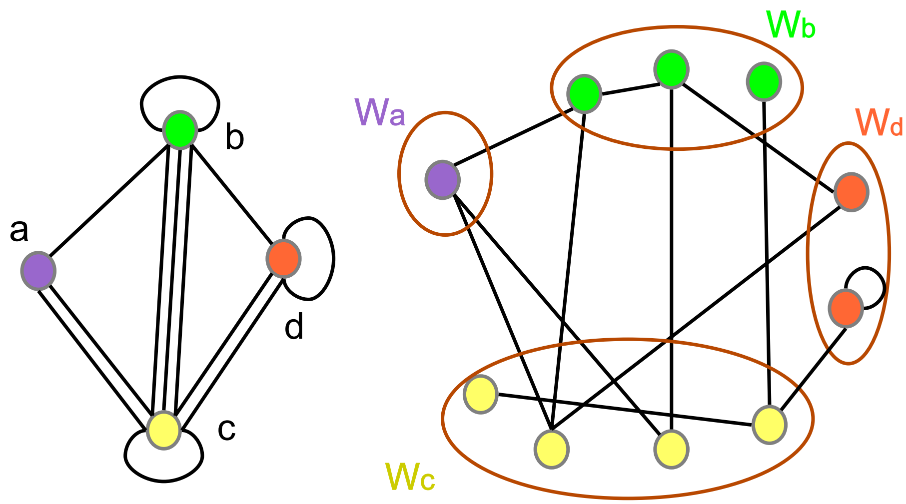
| | |
|---|---|
| H | 4 |
| O | 1 |
| N | 1 |
| C | 3 |
| A | 1 |
| NH | 1 |
| CH | 3 |
| CO | 1 |
| CC | 1 |
| NA | 1 |
| CA | 2 |




$G=(V, E)$

$r(\mathbf{H}) = 4$
$r(\mathbf{O}) = 1$
$r(\mathbf{N}) = 1$
$r(\mathbf{C}) = 3$
$r(\mathbf{A}) = 1$

## Detachment : A Reverse Operation of Contraction




A multigraph *G*
$G = (V, E)$

A detachment *H* of *G*
$H = (\cup_{v \in V} W_v, E)$