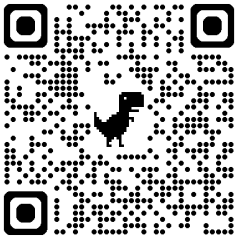
高さ定数の根付き木に対する 極大/飽和頻出部分木マイニング問題の計算複雑性

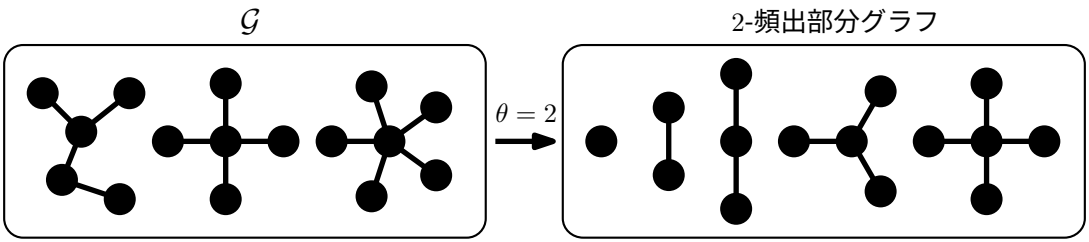
名古屋大学

甲本健太，栗田和宏，小野廣隆



はじめに

頻出部分グラフ: グラフの集合に，ある回数以上部分グラフとして含まれるようなグラフ



頻出部分グラフ列挙問題:
グラフの集合が与えられるとき，その頻出部分グラフを列挙する問題

応用:

- 化学物質のデータベースからのデータ収集
- XMLなどの形式で表現された文書の分析
- RNA配列からのパターン抽出

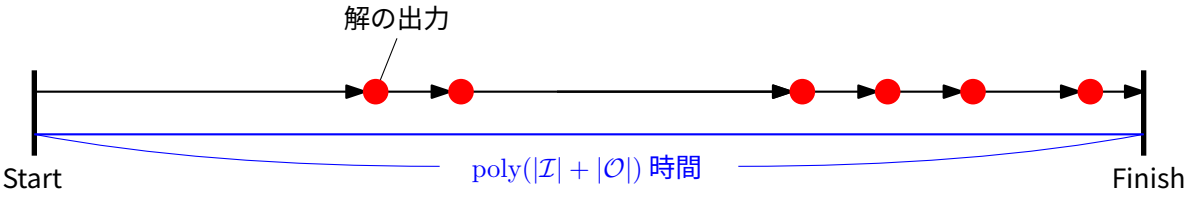
列挙アルゴリズムの計算量評価指標 (1/2)

列挙アルゴリズムでは一般に，出力サイズが入力サイズの多項式で**抑えられない**．
→ ここでは，出力サイズに依存した計算量評価指標を用いる．

定義: 出力多項式時間アルゴリズム

列挙アルゴリズム A が解をすべて出力するまでにかかる時間が，入力サイズと出力サイズの多項式で抑えられるとき，A を**出力多項式時間アルゴリズム**という．

↓ 出力多項式時間アルゴリズムのイメージ



（入力: $\mathcal{I}$，出力: $\mathcal{O}$）

列挙アルゴリズムの計算量評価指標 (2/2)

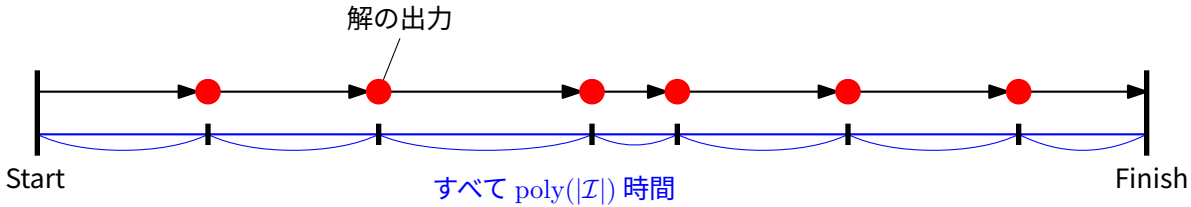
定義: 多項式遅延アルゴリズム

列挙アルゴリズム A の**遅延**とは，A が

1. 最初の解を出力するまでの時間，
2. i 番目の解を出力してから $i+1$ 番目の解を出力するまでの時間，
3. 最後の解を出力してからアルゴリズムが停止するまでの時間，

の最大値である．また，A の遅延が入力サイズの多項式で抑えられるとき，
A を**多項式遅延アルゴリズム**という．

↓ 多項式遅延アルゴリズムのイメージ



定義より，「多項式遅延アルゴリズム $\Rightarrow$ 出力多項式時間アルゴリズム」が言える．

頻出部分グラフ列挙問題 (1/3)

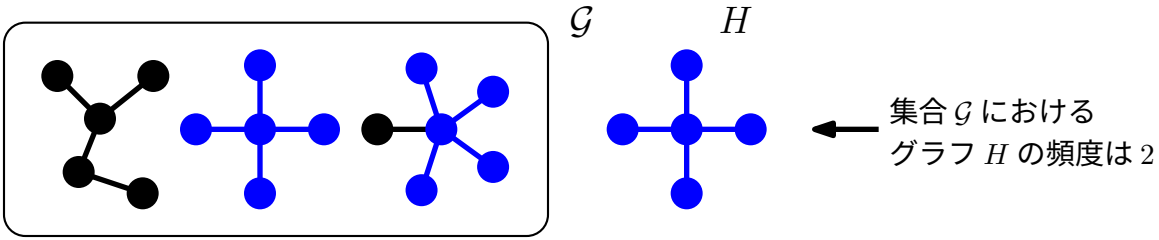
以後，グラフは連結であることを仮定する．
（非連結なグラフの部分グラフ同型性判定は**パスフォレストでも** NP困難）

定義: 頻度

グラフの集合 $\mathcal{G}$ とグラフ H について，グラフ $G \in \mathcal{G}$ で，
H を部分グラフとして含むグラフの個数を $\mathcal{G}$ における H の**頻度**という．

定義: 頻出部分グラフ・共通グラフ

グラフの集合 $\mathcal{G}$ とグラフ H について，$\mathcal{G}$ における頻度が θ 以上であるようなグラフを
$\mathcal{G}$ の θ-**頻出部分グラフ**という．
特に，$\theta = |\mathcal{G}|$ である場合，**共通グラフ**という．

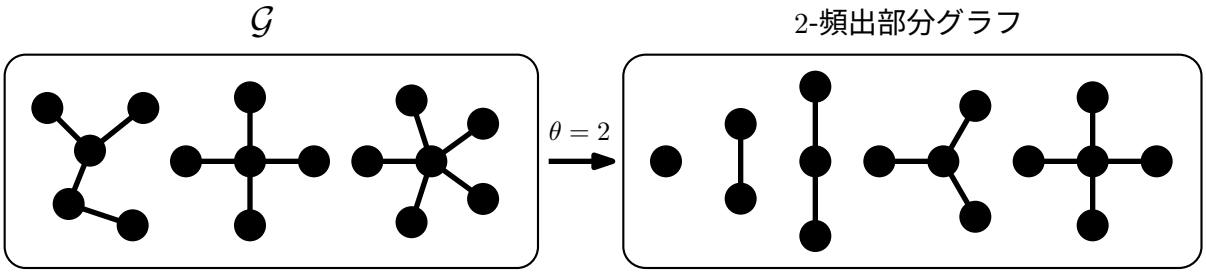


頻出部分グラフ列挙問題 (2/3)

頻出部分グラフ列挙問題は次のように定義される．

定義: 頻出部分グラフ列挙問題

入力：グラフの集合 $\mathcal{G}$，しきい値 $\theta > 0$．
出力：$\mathcal{G}$ の θ-頻出部分グラフすべてからなる集合．



Horváth らにより，木幅有界なグラフの集合に対する頻出部分グラフ列挙問題は
出力多項式時間で解けることが示されている [1]．

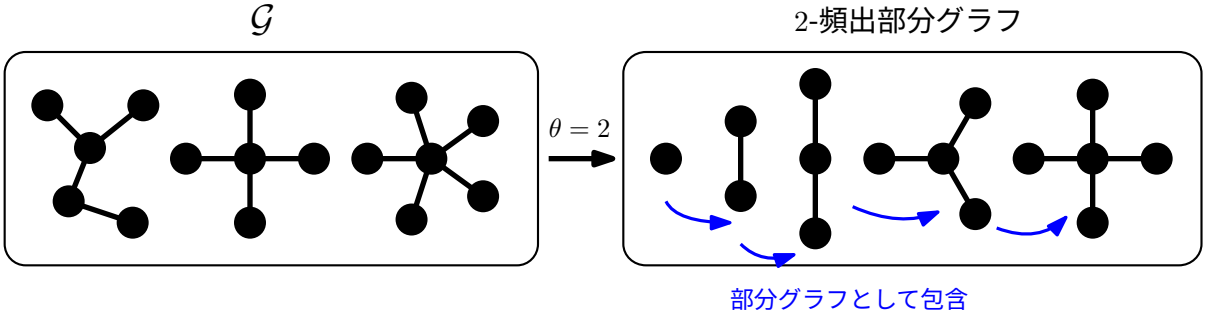
---

[1] Horváth, T., Otaki, K., Ramon, J. Efficient Frequent Connected Induced Subgraph Mining in Graphs of Bounded Tree-Width. ECML PKDD, 2013. https://doi.org/10.1007/978-3-642-40988-2_40

頻出部分グラフ列挙問題 (3/3)

頻出部分グラフ列挙問題の課題

- 頻出部分グラフの数は非常に多くなることがある
- 他の頻出部分グラフに包含されているグラフが出力される場合がある



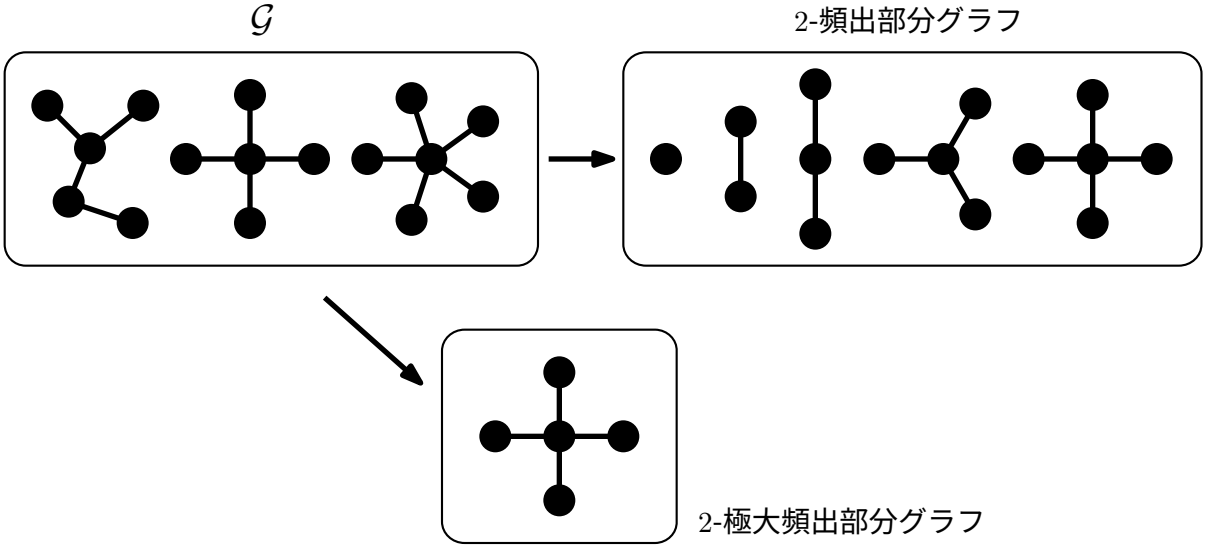
→ 頻出部分グラフのうち，よい性質を満たすものだけ出力したい

ここでは，**極大性**，**飽和性** といった性質を用いる

極大 / 飽和 頻出部分グラフ列挙問題 (1/4)

定義: 極大頻出部分グラフ

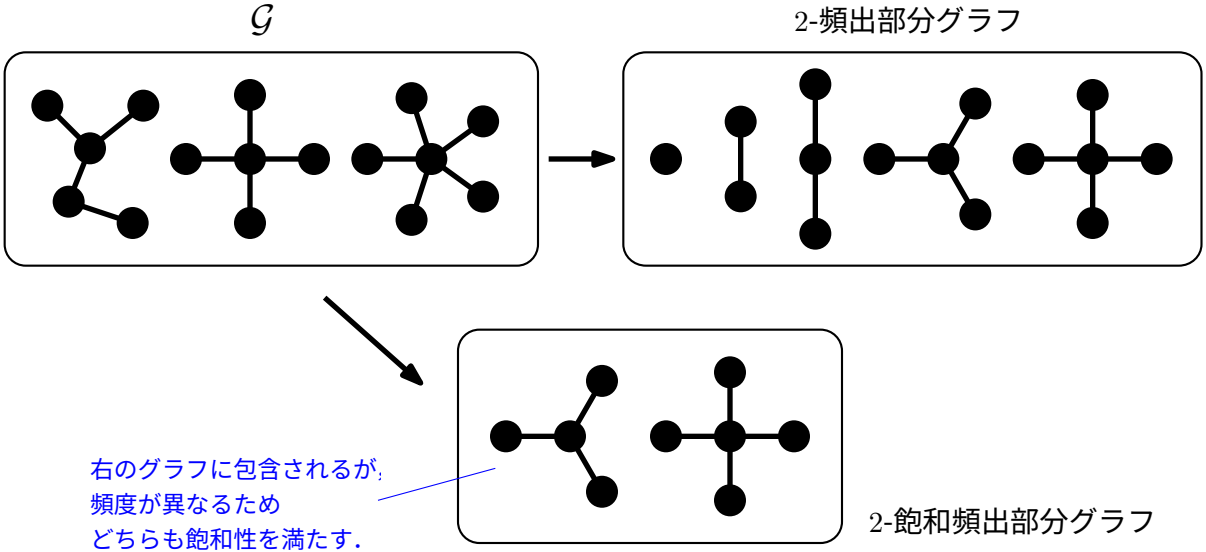
グラフの集合 $\mathcal{G}$ と整数 $\theta > 0$，$\mathcal{G}$ の θ-頻出部分グラフ H について，
$\mathcal{G}$ の H と異なる任意の θ-頻出部分グラフ H' が，H を部分グラフとして真に含まない
とき，H を $\mathcal{G}$ の θ-**極大頻出部分グラフ**という．



極大 / 飽和 頻出部分グラフ列挙問題 (2/4)

定義: 飽和部分グラフ・飽和頻出部分グラフ

グラフの集合 $\mathcal{G}$ とグラフ H について，H を部分グラフとして真に包含し，かつ H と頻度が等しいグラフが存在しないとき H は $\mathcal{G}$ の**飽和部分グラフ**であるという．
さらに，頻度 θ 以上の飽和部分グラフを θ-**飽和頻出部分グラフ**という．



極大 / 飽和 頻出部分グラフ列挙問題 (3/4)

定義: 極大頻出部分グラフ列挙問題・飽和頻出部分グラフ列挙問題

入力: グラフの集合 $\mathcal{G}$，しきい値 $\theta > 0$
出力: $\mathcal{G}$ のすべてのθ-極大 / 飽和 頻出部分グラフからなる集合

Kimelfeld らにより，極大 / 飽和 頻出部分グラフ列挙問題は**入力を木に限定しても**
P = NP でない限り，出力多項式時間で解けないことが示されている [2].

この困難性の証明は高さ 60 程度の木を用いて行われているため，
さらに高さが小さい場合の困難性は明らかになっていない.

→ 本研究では，入力を木に限定し，その高さを制限したときの
極大 / 飽和 頻出部分グラフ列挙問題について考えた.

---

[2] Kimelfeld and Kolaitis. (2013). The complexity of mining maximal frequent subgraphs. PODS 13.
https://doi.org/10.1145/2463664.2465222

極大 / 飽和 頻出部分グラフ列挙問題 (4/4)

高さ h の木の集合に対する頻出部分木マイニングについては，
以下のような結果が知られている．（赤字部分が本研究の結果）

	頻出	極大頻出	飽和頻出
$h = 1$	多項式時間	多項式時間	多項式時間
$h = 2$	出力多項式時間 [1]	**多項式時間 ($\theta = \|\mathcal{G}\|$)**	**多項式遅延**
$h = 3$	出力多項式時間 [1]	?	?
$\vdots$			
$h \geq 60$	出力多項式時間 [1]	NP-hard [2]	NP-hard [2]

ここで，NP-hard とは，P $\neq$ NP の仮定のもとで
出力多項式時間アルゴリズムが存在しないことをいう．

---

[1] Horváth, T., Otaki, K., Ramon, J. Efficient Frequent Connected Induced Subgraph Mining in Graphs of Bounded Tree-Width. ECML PKDD, 2013. https://doi.org/10.1007/978-3-642-40988-2_40

[2] Kimelfeld and Kolaitis. (2013). The complexity of mining maximal frequent subgraphs. PODS 13. https://doi.org/10.1145/2463664.2465222

結果

本研究では，以下の定理を示した．

定理 1

高さ 2 以下の根付き木の集合 $\mathcal{T}$ に対して $\mathcal{T}$ の部分グラフ同型性についての極大性に関して極大な共通木は一意に定まり，多項式時間で求められる．

定理 2

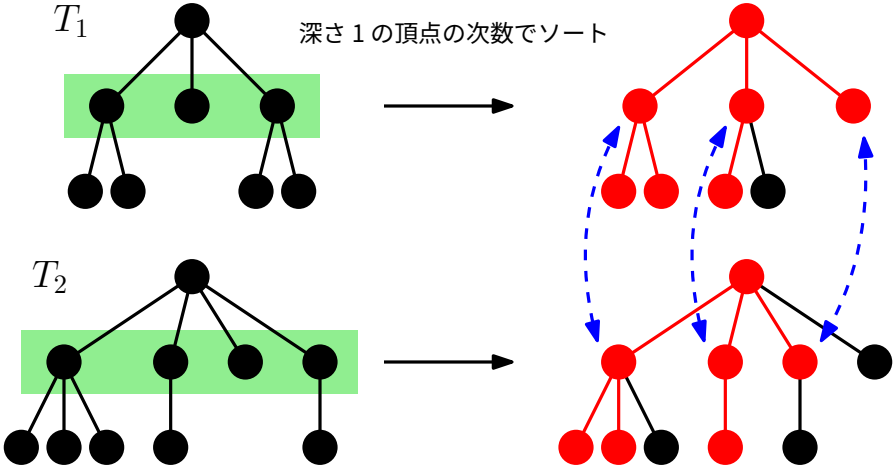
高さ 2 以下の根付き木の集合に対し，飽和頻出部分木は多項式遅延で列挙できる．

本発表では，定理 1, 2 について述べる．

高さ 2 の根付き木の極大共通木

木が 2 つの場合

高さ 2 の根付き木 T_1, T_2 を考える．T_1, T_2 の深さ 1 の頂点を次数が大きい順にとり，
その部分木について極大な共通木をとる．→ 極大共通木となり，ただ一つに定まる．



木が n 個の場合

木が n 個の場合でも，極大な共通木はただ一つに定まり，多項式時間で求まる．
$\mathcal{T} = \{T_1, \ldots, T_n\}$ の極大な共通木を $\mathrm{MCT}(\mathcal{T})$ と表す．

飽和頻出部分木の列挙 (1/8)

先の結果（定理 1）を用いて，下記の定理 2 を示す．

定理 2

高さ 2 以下の根付き木の集合に対し，飽和頻出部分木は多項式遅延で列挙できる．

飽和頻出部分木の列挙 (2/8)

まず，木 X の**サポート**を次のように定義する．

定義: サポート

木の集合 $\mathcal{T}$ と木 X に対して，$\mathcal{T}$ 中の木うち X を部分グラフとして含むグラフの集合を $\mathcal{T}$ における X の**サポート** といい，$\mathcal{T}(X)$ と表す．

このとき，飽和頻出部分木の定義から以下の補題が成り立つ．

補題

X が飽和部分木である $\Leftrightarrow \mathrm{MCT}(\mathcal{T}(X)) = X$．

また，定理1より $\mathcal{T}$ の部分集合 S を固定したとき，S の極大な共通木はただ一つに定まる．

よって，$\mathrm{MCT}(\mathcal{T}(X)) = X$ を満たす木 X を列挙できればよい．

逆探索と呼ばれる手法を用いてこのような木を列挙する．

飽和頻出部分木の列挙 (3/8)

逆探索

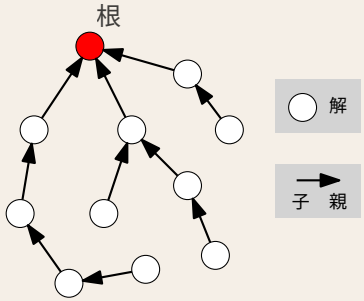
列挙木（下記参照）を深さ優先探索して解を列挙する手法

定義: 列挙木

列挙問題の解の集合 $\mathcal{S}$ に，

- **根**と呼ばれる特別な解 R を定義する．
- 根を除く全ての解 S に対して，ちょうど 1 つ，**親**となる解 $p(S)$ を定義する．

ここで，根を除く全ての解から，その親に辺を張った根付き木を**列挙木**という．



逆探索のうれしさ

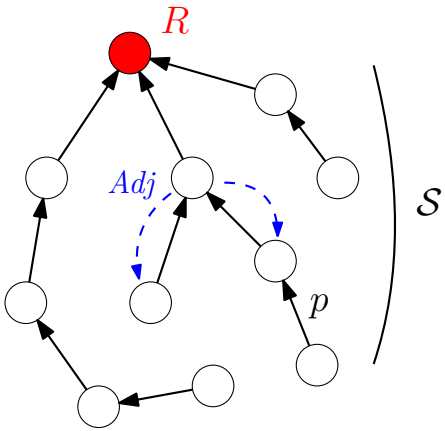
- **根の計算**，**親の計算**，i **番目の子の計算**がそれぞれ入力サイズの多項式時間でできると，解の列挙にかかる時間を抑えられる．
- 出力済みの解を記憶しておく必要がないため，空間計算量が小さい．

飽和頻出部分木の列挙 (4/8)

列挙木の構築

列挙木は明に構築する必要はなく，
解の集合を $\mathcal{S}$ として，以下を定義すればよい:

- **根** $R \in \mathcal{S}$
- **親を求める関数** p
- **子を求める関数** Adj



ただし，以下の条件を満たす必要がある:

1. 親を有限回たどることで，根に到達できる．
2. 親の子に自身が含まれる．

飽和頻出部分木の列挙 (5/8)

列挙木の定義

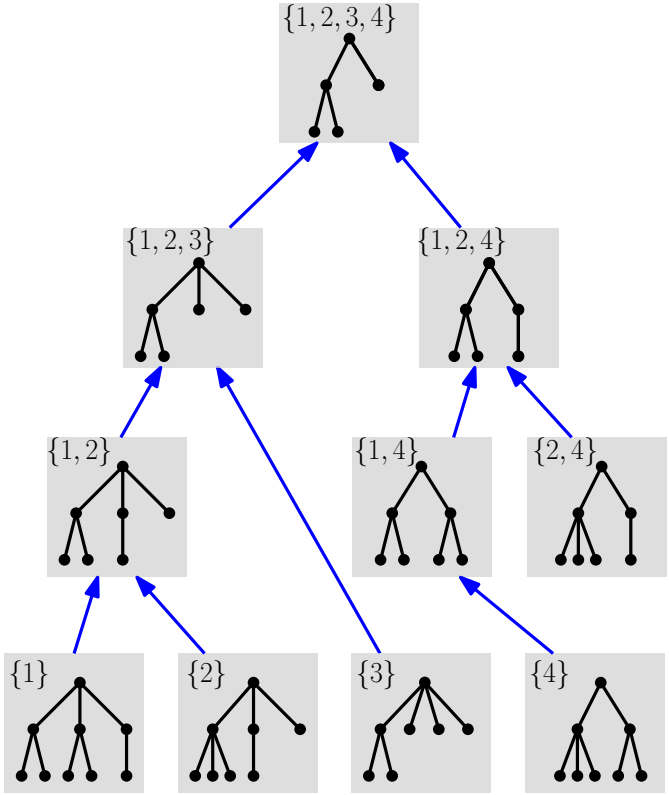
根の計算:
$\mathrm{MCT}(\mathcal{T})$ を根とする．

親の計算:
$\mathcal{T}(X)$ に属していない $\mathcal{T}$ の要素で，
$\mathrm{MCT}(\mathcal{T}(X) \cup \{X'\})$ が極小となる
木 X' に対して，

$$p(X) := \mathrm{MCT}(\mathcal{T}(X) \cup \{X'\}).$$

子の計算:
木 X に次数 1 の頂点を 1 つ加えた木
X' に対して，$\mathrm{MCT}(\mathcal{T}(X'))$．

上の集合はその木のサポートを表す →



飽和頻出部分木の列挙 (6/8)

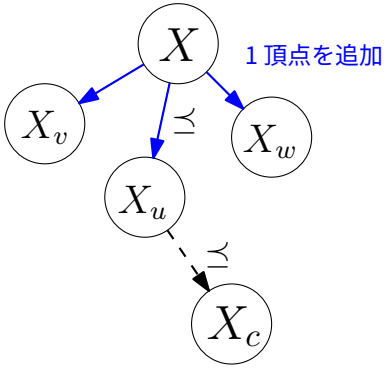
列挙木が満たすべき条件の確認

1. 親を有限回たどることで根に到達できる

任意の飽和部分木 X とその親 $p(X)$ に対して，$|\mathcal{T}(X)| < |\mathcal{T}(p(X))|$.
→ 1 回親をたどることでそのサポートは少なくとも 1 以上増える.
→ 高々 $|\mathcal{T}|$ 回親をたどれば根に到達.

2. 親の子に自身が含まれている

任意の木 X_c と X_c の親 X について考える.
X に 1 頂点追加して得られる木で X_c に包含される木が存在する. （右図 X_u）
ここで，$\mathrm{MCT}(\mathcal{T}(X_u)) \neq X_c$ と仮定すると，
親の極小性に矛盾.

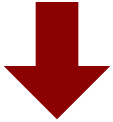


飽和頻出部分木の列挙 (7/8)

計算量の解析

逆探索にかかる時間計算量は，$O(t(R) + (t(Adj) + t(p)) \cdot \Delta_{\mathcal{F}} \cdot h_{\mathcal{F}})$ 遅延．

- 任意の解から根までは，高々 $|\mathcal{T}|$ 回親を辿れば到達可能 → $h_{\mathcal{F}} \leq |\mathcal{T}|$
- 子の数も高々 $|\mathcal{T}|$ → $\Delta_{\mathcal{F}} \leq |\mathcal{T}|$



列挙アルゴリズムの遅延は，
入力サイズの多項式で抑えられる．

よって，
飽和部分技を多項式遅延で列挙できた．

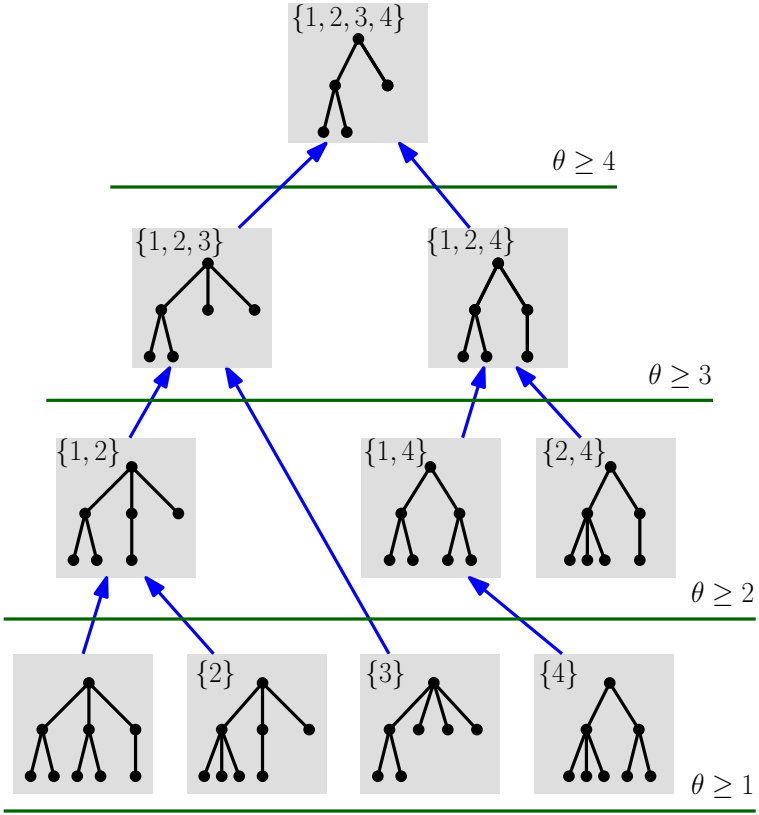
変数	値
$t(R)$	根の計算にかかる時間
$t(Adj)$	i 番目の子の計算にかかる時間
$t(p)$	親の計算にかかる時間
$\Delta_{\mathcal{F}}$	列挙木 $\mathcal{F}$ の次数の最大値
$h_{\mathcal{F}}$	列挙木 $\mathcal{F}$ の高さ

飽和頻出部分木の列挙 (8/8)

頻度 θ 以上の解の列挙

この列挙木は，根のサポート $|\mathcal{T}|$ を最大値として，葉に近づくほど単調にサポートが減少していく．

θ-飽和頻出部分木を列挙したい場合，サポート θ の解から先の探索を打ち切れば良い．

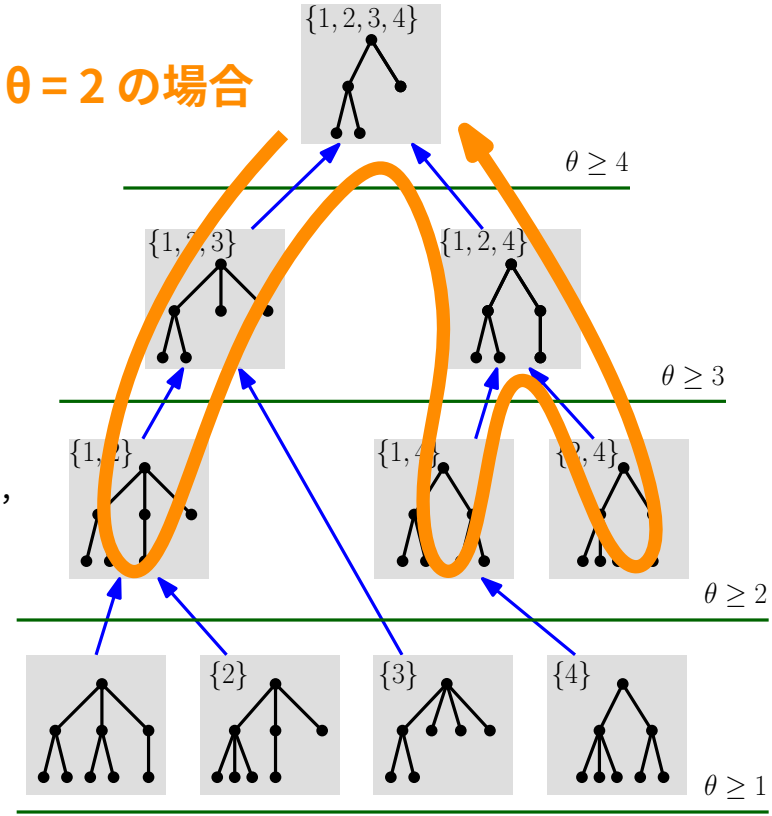


飽和頻出部分木の列挙 (8/8)

頻度 θ 以上の解の列挙

この列挙木は，根のサポート $|\mathcal{T}|$ を最大値として，葉に近づくほど単調にサポートが減少していく．

θ-飽和頻出部分木を列挙したい場合，サポート θ の解から先の探索を打ち切れば良い．



まとめと今後の展望

本研究の結果

- 高さ 2 以下の根付き木の集合に対して極大な共通木は一意に定まり，
  またその木は多項式時間で求められる．
- 高さ 2 以下の根付き木に対する飽和頻出グラフは多項式遅延で列挙できる．

今後の展望

- 飽和頻出部分木マイニングが計算困難になる高さの境界
  - 高さ 60 以上の木では計算困難であることが示されている [2]．
  - 本研究で高さ 2 以下の木では多項式遅延アルゴリズムが存在することを示した．
    → よって，境界となる高さ h は $3 \leq h \leq 60$
- 高さ 2 以下の極大頻出部分木マイニングを行うアルゴリズム

---

[2] Kimelfeld and Kolaitis. (2013). The complexity of mining maximal frequent subgraphs. PODS 13.
https://doi.org/10.1145/2463664.2465222