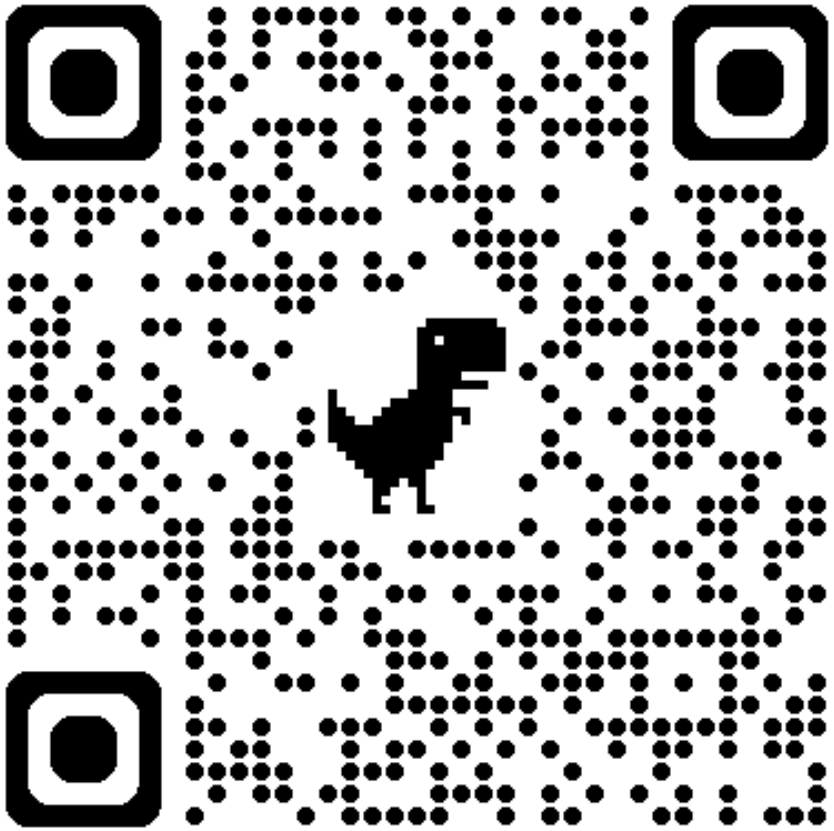
ラベル付き木に対する極大頻出部分木マイニングの計算複雑性

名古屋大学

甲本健太, 栗田和宏, 小野廣隆

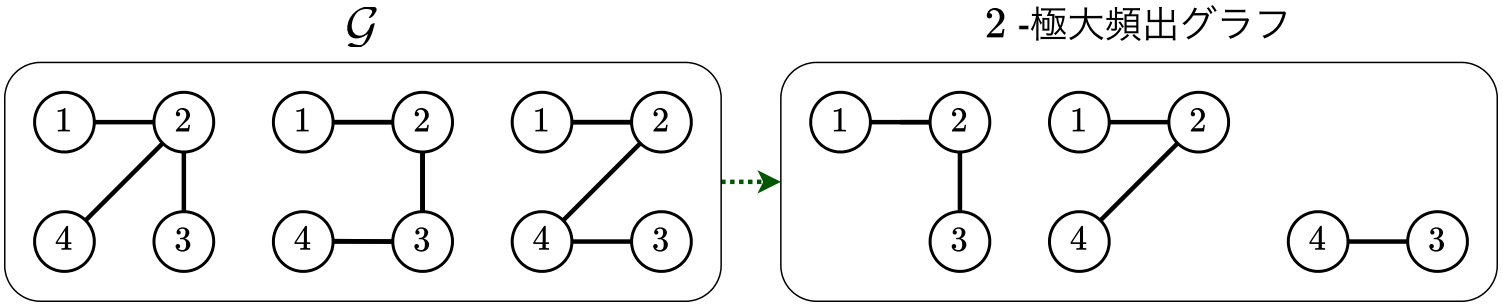
スライド情報



kentakom1213.github.io/slides/2025-enum-seminar.pdf

研究の全体像 (1/2)

頻出グラフとは，グラフの集合に繰り返し現れるような部分グラフである．
また，頻出グラフが他の頻出グラフに部分グラフとして含まれていないとき，
極大頻出グラフという．



グラフの集合が与えられたとき，その極大頻出グラフを列挙する問題は
極大頻出グラフマイニング問題と呼ばれ，以下のような応用が知られる．

- 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

研究の全体像 (2/2)

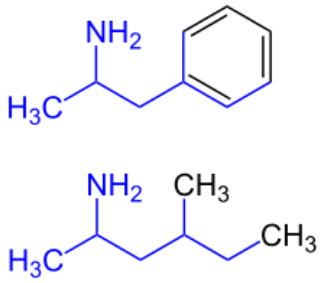
化学物質のデータベースからのデータ収集の例

- 分子構造は，原子をラベルとしたラベル付きグラフとしてモデル化される
- 分子構造が近い化合物は互いに似た性質を持つことがある

→ 頻出グラフマイニングにより化合物のデータベースから効率的にデータ収集可能

上がアンフェタミン，下がメチルヘキサンアミン，
どちらも神経を興奮させる作用を持つ．

画像: https://en.wikipedia.org/wiki/Chemical_similarity



さらに，分子構造は次数が小さく，木幅が小さいグラフであることが多い
→ 入力を限定した効率的なアルゴリズムを設計できる場合がある

本研究では，入力を木の集合に制限したときの極大頻出グラフマイニング問題の計算複雑性について考察した．

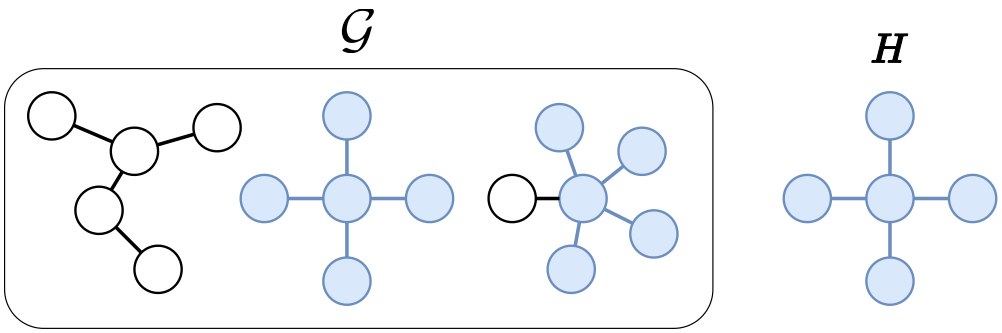
準備: 頻出グラフ

定義: 頻度

グラフの集合 $\mathcal{G}$ とグラフ H が与えられたとき，$\mathcal{G}$ の要素で，H を部分グラフとして含むようなグラフの個数を $\mathcal{G}$ における H の**頻度**という．

定義: 頻出グラフ

グラフの集合 $\mathcal{G}$ と整数 $t > 0$ が与えられたとき，$\mathcal{G}$ における頻度が t 以上であるようなグラフを $\mathcal{G}$ の **t-頻出グラフ**という．また，t が明らかな場合は単に**頻出グラフ**という．特に $t = |\mathcal{G}|$ である場合，**共通グラフ**という．



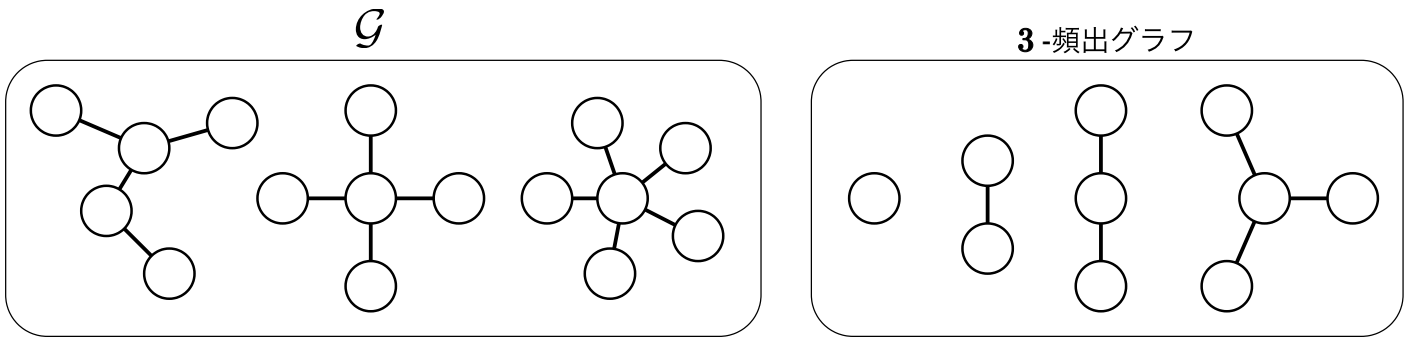
グラフの集合 $\mathcal{G}$ における
グラフ H の頻度は 2

頻出グラフ列挙問題 (1/2)

極大でない頻出グラフマイニング問題は以下のように定義される.

頻出グラフ列挙問題 (FCISM)

入力: グラフの集合 $\mathcal{G}$ と整数 $t > 0$
出力: $\mathcal{G}$ のすべての t-頻出グラフからなる集合 $\mathcal{F}_t$



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

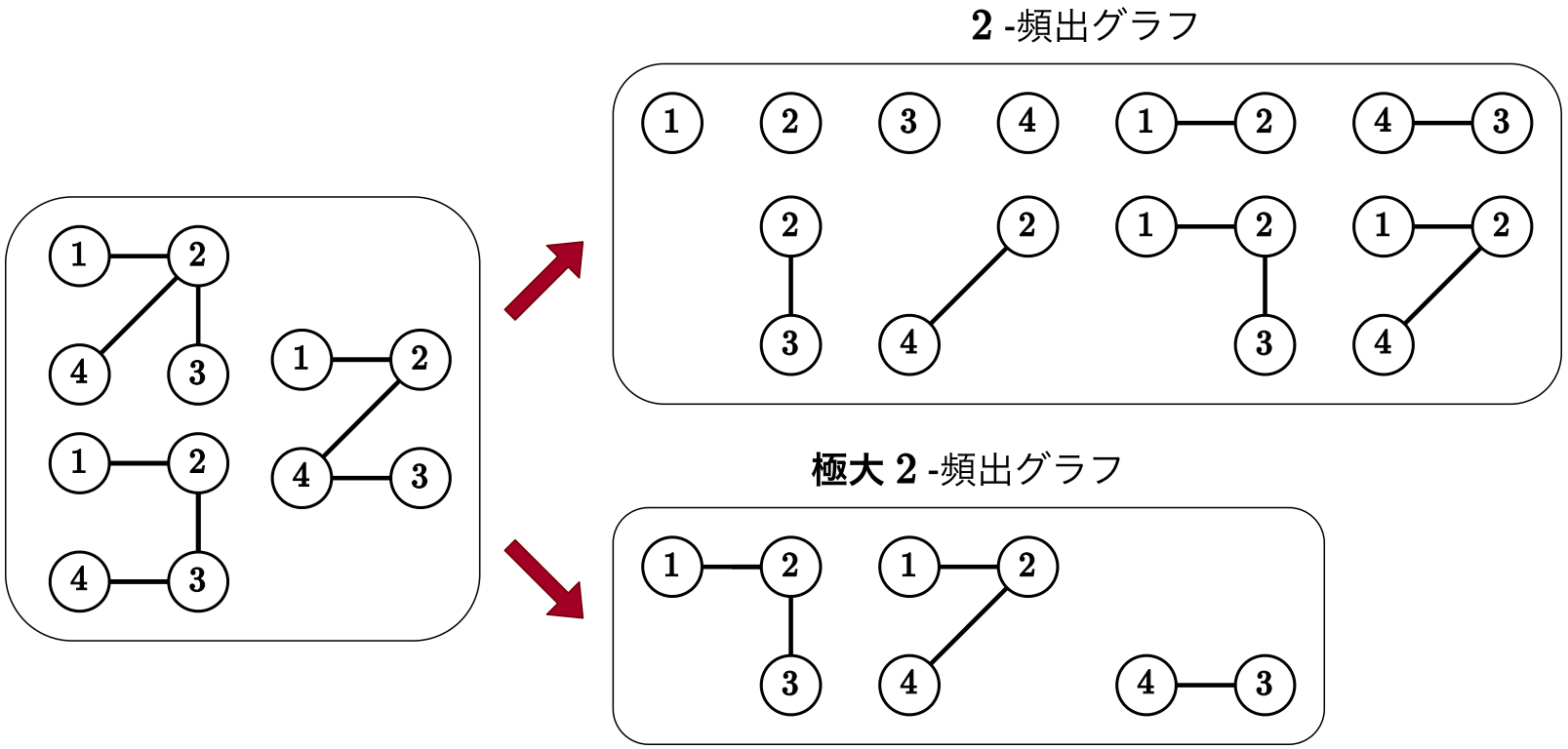
---

[1] Horváth, T., Otaki, K., Ramon, J. (2013). 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

頻出グラフ列挙問題 (2/2)

頻出グラフ列挙問題の課題

頻出グラフの数は非常に多くなることがあり，あまり有用ではない情報も含まれる
→ 部分グラフについての包含関係に対する**極大性**でフィルタリング



極大頻出グラフ列挙問題

頻出グラフの極大性と極大頻出グラフ列挙問題の定義

定義: 頻出グラフの極大性

グラフの集合 $\mathcal{G}$ の t-頻出グラフ H が，$\mathcal{G}$ の他の t-頻出グラフに部分グラフとして含まれていないとき，頻出グラフ H は**極大**であるという.

極大頻出グラフ列挙問題 (MaxFCISM)

入力: グラフの集合 $\mathcal{G}$ と整数 $t > 0$
出力: $\mathcal{G}$ のすべての**極大な** t-頻出グラフからなる集合 $\mathcal{F}_t$

Kimelfeld らにより，入力を 2 つの木に限定しても MaxFCISM 問題は計算困難であることが示されている [2].

本研究では，入力を限定したときの MaxFCISM 問題の計算複雑性を調べた.

---

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

結果

以下の 2 つの定理を示した.

定理 1

入力を，グラフの集合 $\mathcal{G}$ に含まれるグラフが**ラベル付きのスター**である場合に限定しても，MaxFCISM 問題は計算困難である.

定理 2

入力を，$\mathcal{G}$ に含まれるグラフがすべて**高さ 2 のラベルなし木**かつ，$t = |\mathcal{G}|$ である場合に限定したとき，$\mathcal{G}$ の共通木は一意に定まり，MaxFCISM 問題には多項式時間アルゴリズムが存在する.

本発表では，定理 1，定理 2 の証明の概略を説明する.

準備: 列挙アルゴリズムの計算複雑性 (1/2)

列挙アルゴリズムの計算複雑性クラス

定義

列挙アルゴリズム A が**出力多項式時間**で動作するとは，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ことをいう.

本発表では，列挙問題 X に出力多項式時間アルゴリズムが $\mathrm{P} = \mathrm{NP}$ でない限り存在しないとき，列挙問題 X は**計算困難**であるという.

準備: 列挙アルゴリズムの計算複雑性 (2/2)

列挙問題の計算困難性の証明法

- 本研究では，列挙問題を**別解問題**と呼ばれる決定問題に言い換えて議論する.
- 別解問題が NP-困難であれば，もとの列挙問題も計算困難である.

別解問題とは？

- 列挙問題
  - 問題の解をすべて出力する
- 別解問題
  - 問題と，解の集合 $\mathcal{S}$ が与えられる
    → $\mathcal{S}$ に含まれていない解が存在するか？（Yes / No）

ラベル付きスターの場合の計算困難性 (1/4)

定理 1 の証明の概略

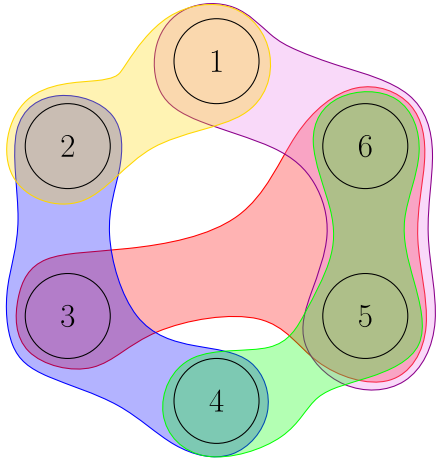
極大頻出アイテム集合列挙問題の計算困難性から帰着する.

ハイパーグラフ

- グラフの一般化
- 頂点集合 V とハイパーエッジ E $(\subseteq V)$ の集合 $\mathcal{E}$ の組

頂点 V を列，ハイパーエッジ $\mathcal{E}$ を行とする，値が $0, 1$ の接続行列を用いて表す.
$\rightarrow n$ 頂点 m 辺のハイパーグラフでは，$m \times n$ 行列

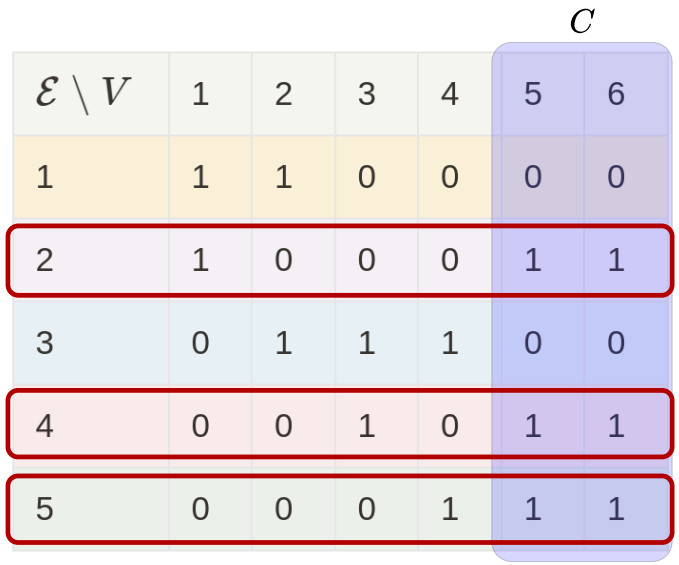
$\mathcal{E} \setminus V$	1	2	3	4	5	6
1	1	1	0	0	0	0
2	1	0	0	0	1	1
3	0	1	1	1	0	0
4	0	0	1	0	1	1
5	0	0	0	1	1	1



ラベル付きスターの場合の計算困難性 (2/4)

定義: 頻出アイテム集合

$m \times n$ の 2 値行列 A について，列の部分集合 C が **t-頻出アイテム集合**であるとは，A に，C に属する全ての要素が 1 である行が t 行以上あることをいう.



$\mathcal{E} \setminus V$	1	2	3	4	5	6
1	1	1	0	0	0	0
2	1	0	0	0	1	1
3	0	1	1	1	0	0
4	0	0	1	0	1	1
5	0	0	0	1	1	1

$C = \{5, 6\}$ と選択すると，
列 $2, 4, 5$ は C に属する要素がすべて 1

↓

C は 3-頻出アイテム集合

頻出アイテム集合についても，包含関係について極大性を定義する.

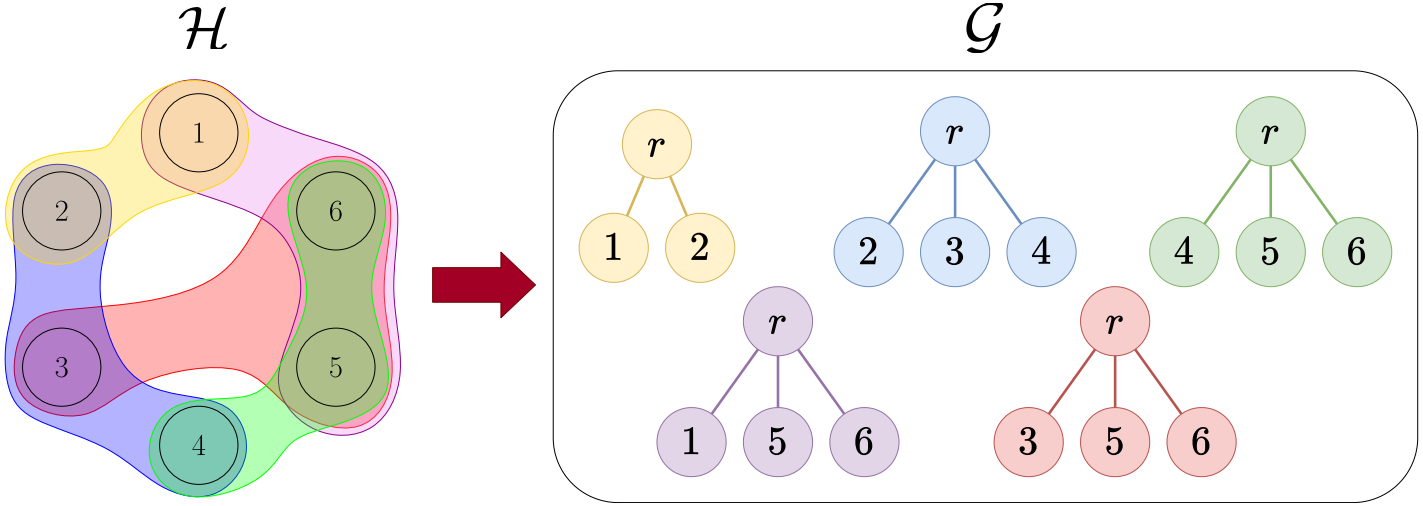
ラベル付きスターの場合の計算困難性 (3/4)

頻出アイテム集合列挙問題 (MaxFIM)

入力: ハイパーグラフ $\mathcal{H} = (V, \mathcal{E})$ と整数 $t > 0$
出力: $\mathcal{H}$ のすべての極大な頻出アイテム集合からなる集合族 $\mathcal{M}_t$

MaxFIM 問題からラベル付きスターの MaxFCISM 問題への帰着

$\mathcal{H}$ の各ハイパー辺 E に対して，根 r と $v \in E$ の間に辺を張ったスターを生成する
$\rightarrow$ 生成したスターの集合を $\mathcal{G}$ とする



ラベル付きスターの場合の計算困難性 (4/4)

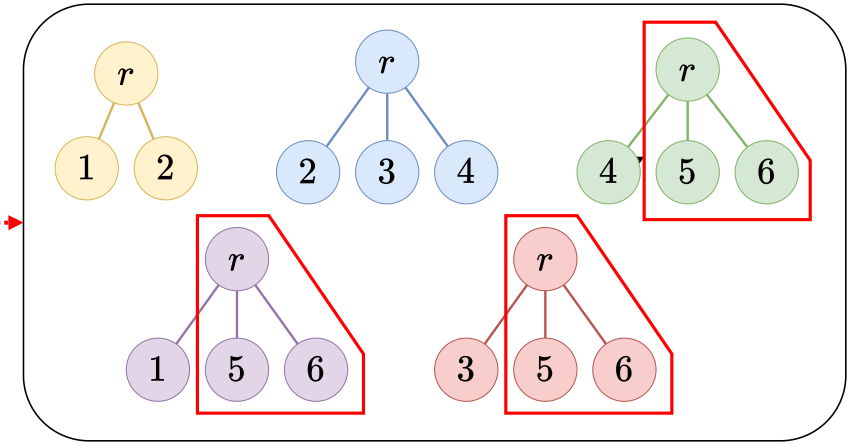
$\mathcal{G}$ と整数 t を入力とした MaxFCISM 問題が解けると仮定する.

$\rightarrow$ 得られた 極大 t-頻出グラフの集合 $\mathcal{F}_t$ から,
極大 t-頻出アイテム集合の集合族 $\mathcal{M}_t$ を復元できる

$\mathcal{H}$

$\mathcal{E} \setminus V$	1	2	3	4	5	6
1	1	1	0	0	0	0
2	1	0	0	0	1	1
3	0	1	1	1	0	0
4	0	0	1	0	1	1
5	0	0	0	1	1	1

$\mathcal{G}$



Boros らにより MaxFIM 問題は計算困難であることが示されている [3].

よって, MaxFIM 問題の計算困難性から MaxFCISM 問題の計算困難性が示される.

---

[3] Boros, Gurvich, Khachiyan, and Makino. On Maximal Frequent and Minimal Infrequent Sets in Binary Matrices. Annals of Mathematics and Artificial Intelligence 39, 211–221 (2003). https://doi.org/10.1023/A:1024605820527

高さ 2 の木に対する共通木マイニング (1/3)

定理 2

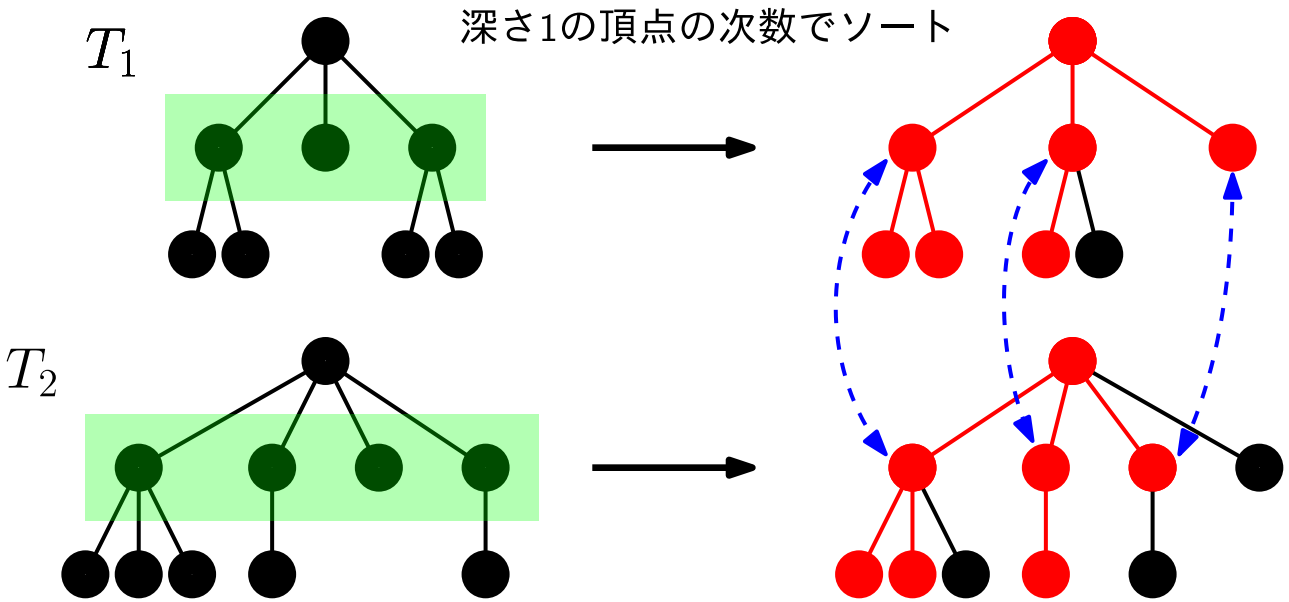
入力を，$\mathcal{G}$ に含まれるグラフがすべて**高さ 2 のラベルなし木**かつ，$t = |\mathcal{G}|$ である場合に限定したとき，$\mathcal{G}$ の共通木は一意に定まり，MaxFCISM 問題には多項式時間アルゴリズムが存在する.

次ページ以降では，定理 2 の証明の概略を説明する.

定理 1 とは異なり，グラフが**ラベルを持たない**場合を考える.

高さ 2 の木に対する共通木マイニング (2/3)

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．$\rightarrow$ 極大共通木はただ一つに定まる．

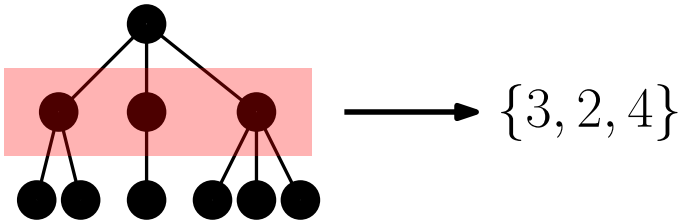


n 個の高さ 2 の根付き木の場合でも，その極大な共通木はただ一つに定まる．
$\rightarrow$ 共通木は多項式時間で出力可能

高さ 2 の木に対する共通木マイニング (3/3)

具体的な手法:

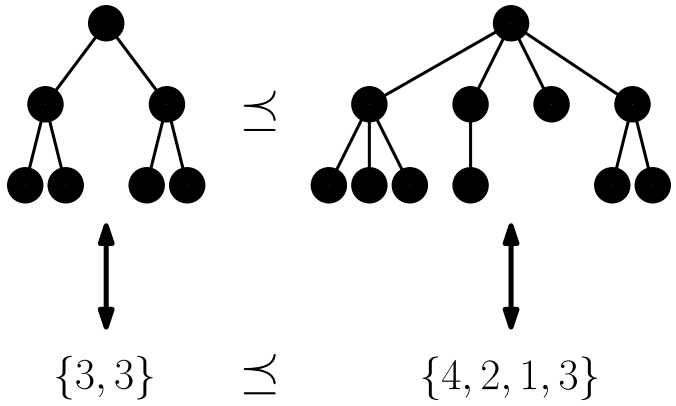
高さ 2 以下の木 T を,
T **の深さ 1 の頂点の次数の多重集合**に変換する.
（1 対 1 に対応する）



定義: 正整数の多重集合同士の包含関係 ($\preceq$)

正整数の多重集合 A, B に対して，単射 $\varphi : A \to B$ が存在し，任意の $x \in A$ について $x \leq \varphi(x)$ であるとき，A は B に包含されるという.

このとき，多重集合同士の包含関係は,
グラフの部分グラフ同型性に対する包含関係
に対応する.



まとめと今後の展望

今回わかったこと

入力を限定した MaxFCISM 問題について考え，以下の 2 つの定理を示した.

1. 入力を，グラフの集合 $\mathcal{G}$ に含まれるグラフが**ラベル付きのスター**である場合に限定しても，計算困難である.
2. 入力を，$\mathcal{G}$ に含まれるグラフがすべて**高さ 2 のラベルなし木**であり，$t = |\mathcal{G}|$ である場合に限定したとき，多項式時間アルゴリズムが存在する.

今後調べたいこと

定理 2 の設定について,

- 高さ h 以下の木の集合に対する共通木マイニングが計算困難になる h は？
- 高さ 2 の場合，t を変えたときの困難性は？

付録: 別解問題の NP-困難性から列挙問題の困難性への帰着

「列挙問題を出力多項式時間で解ける $\Rightarrow$ 別解問題を多項式時間で解ける」を示す.

別解問題のインスタンス: グラフの集合 $\mathcal{G}$, 既知の頻出グラフの集合 $\mathcal{F}_t$

Maximal-FCISM を解く出力多項式時間アルゴリズムを A とする.
$\rightarrow$ ある多項式 f が存在して, A は出力のサイズが s であるようなサイズ i の入力に対して$f(i, s)$時間で停止.

$T = f(|\mathcal{I}|, |\mathcal{S}|)$ として, アルゴリズム A を入力 $\mathcal{I}$ に対して T 時間動作させる.

- A が時間 T 以内に停止: その出力 $\mathcal{O}$ と $\mathcal{S}$ を比較することで,
  合計で $O(T + |\mathcal{O}||\mathcal{S}|)$ 時間で別解問題を解くことができる.
- A が時間 T 以内に停止しない:
  $\mathcal{F}_t$ のサイズは $\mathcal{S}$ のサイズよりも真に大きいことがわかる.
  $\rightarrow$ よって他の解が存在するので, 別解問題の答えは YES.

いずれの場合にも別解問題を多項式時間で解くことができる.

付録: 2 つの木に対する極大頻出部分木マイニングの困難性

Kimelfeld らによる証明 [2] の概要

右のような形の根付き木で
CNF 式をシミュレートする.

↓

SAT の困難性から帰着

（詳細までは読めていない）

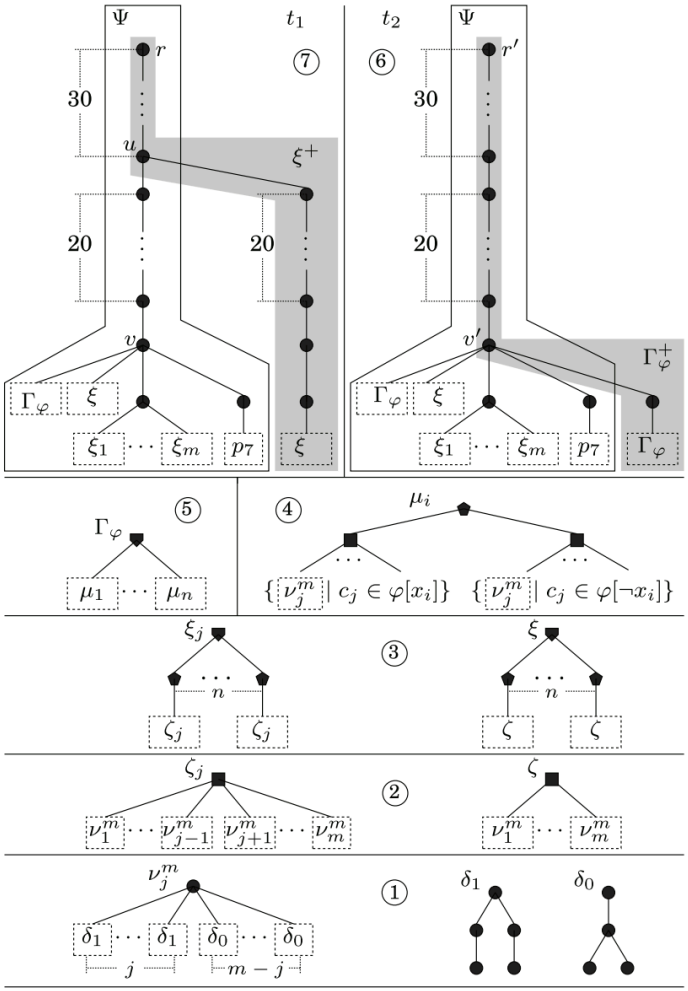


Fig. 5. Constructions in the reduction.

画像: [2] Kimelfeld and Kolaitis. (2013). The complexity of mining maximal frequent subgraphs. PODS '13.