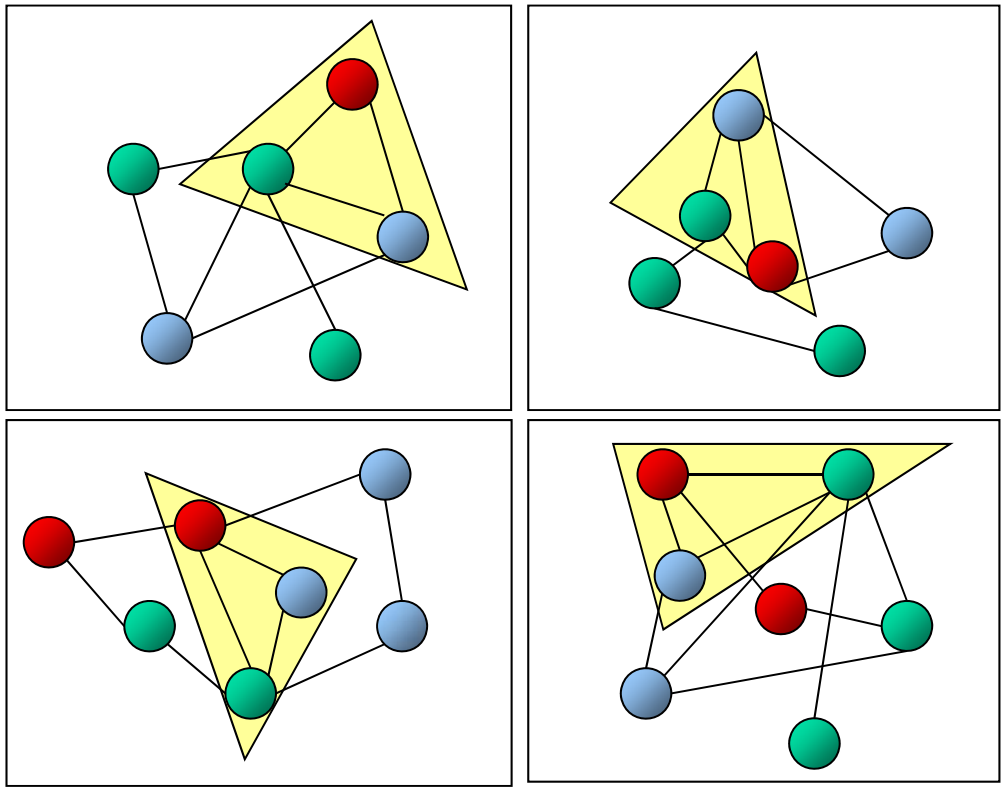
頂点により誘導される頻出グラフ系列パターンのマイニング

猪口 明博，鷲尾 隆
大阪大学 産業科学研究所

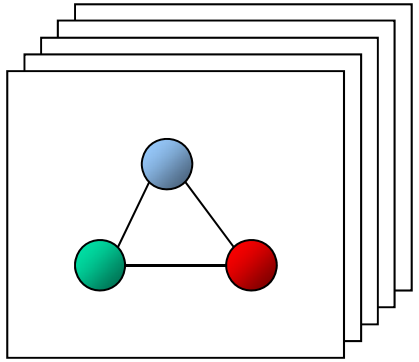
SDM2010にて発表した内容です．

頻出部分グラフマイニング



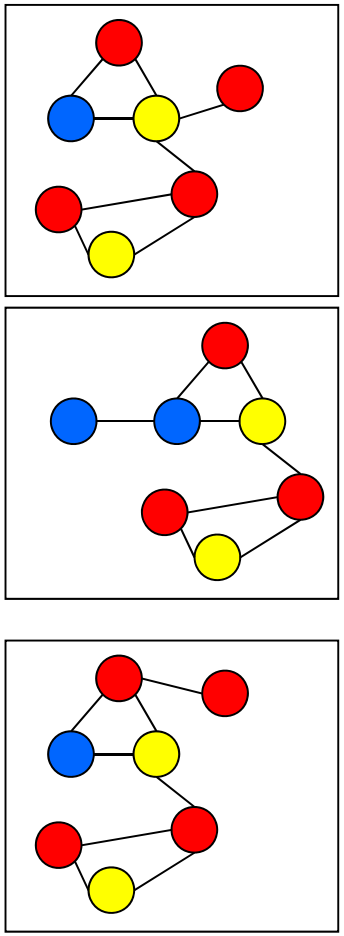
入力

問題:
σ個以上のグラフに含まれる
全ての部分グラフを全て列挙

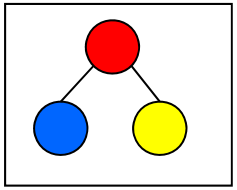


出力

頻出部分グラフマイニング

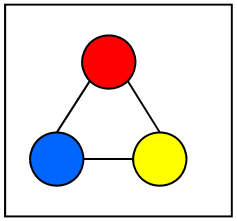


3つのグラフに含まれる
部分グラフのマイニング



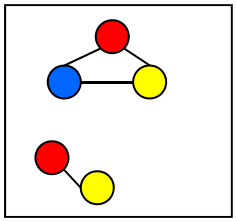
マイニングされたパターンを含む多くのグラフにおいて，赤と黄色の間に辺があるか，ないかは元のグラフデを見ないと分からない（理解困難）

3つのグラフに含まれる
誘導部分グラフのマイニング



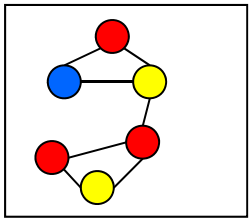
パターンを理解するためにデータベース中のグラフをみる必要がない（理解困難ではない）

3つのグラフに含まれる
誘導部分グラフのマイニング



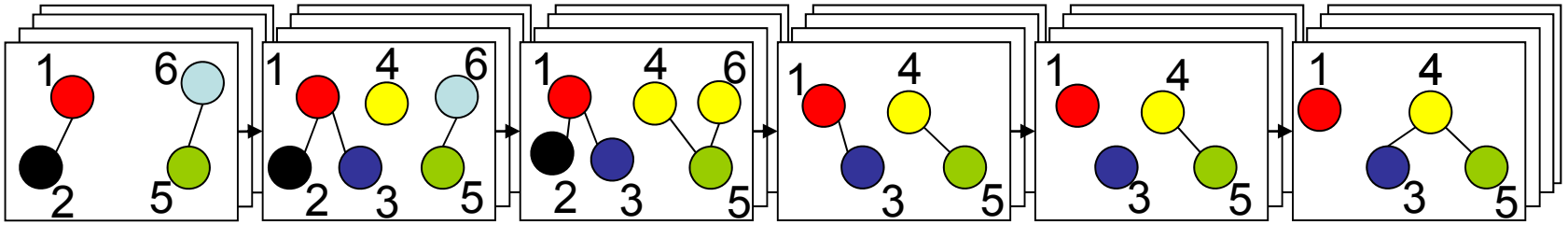
マイニングされたパターンを含む多くのグラフにおいて，上と下の連結部分を直接結ぶ辺がないことは分かるが，それらがどのような関係であったかは元のグラフデータを見ないと分からない（理解困難）

3つのグラフに含まれる
「連結」誘導部分グラフの
マイニング

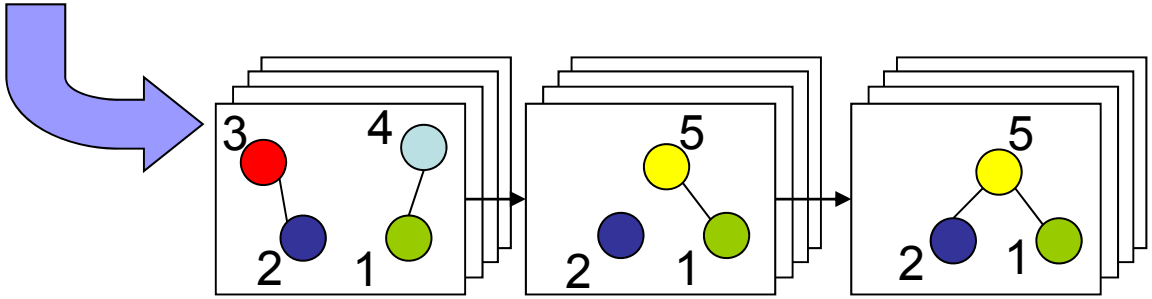


パターンを理解するためにデータベース中のグラフをみる必要がない（理解困難ではない）

グラフ系列のマイニング



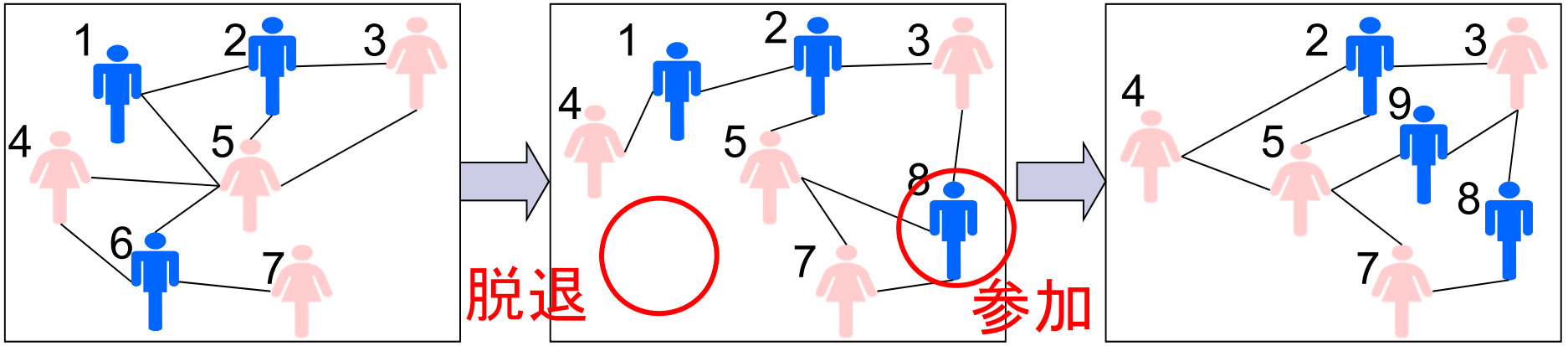
頻出する部分グラフ系列を
マイニング



- グラフ系列
  - 頂点数，辺数が増減する．
  - 頂点ラベル，辺ラベルが変化する．
  - 各頂点は，IDをもつ．

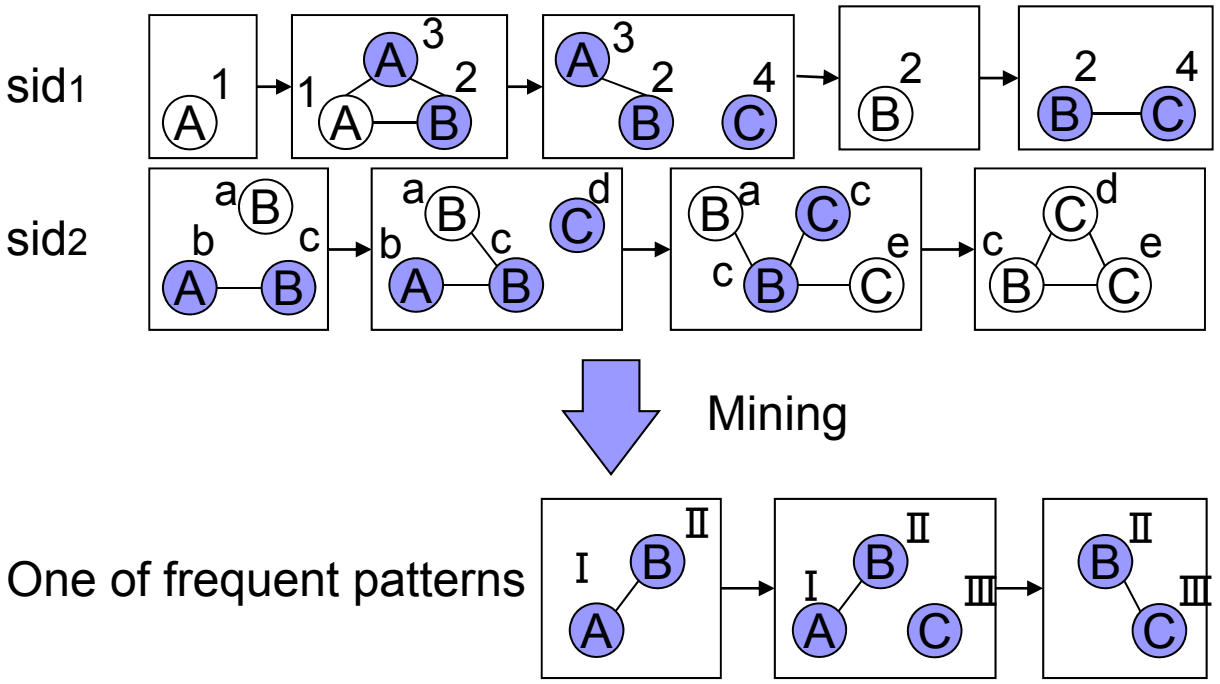
グラフ系列の例

- 人間関係ネットワークの変化
  - 人：頂点，人間関係：辺
- ホームページのリンク構造の変化
  - HTML文章：頂点，ハイパーリンク：辺
- 遺伝子ネットワークの変化(進化)
  - 遺伝子：頂点，相互作用：辺
- 機械の組み立て
  - 部品：頂点，隣接する部品間：辺
- その他．．．



問題定義

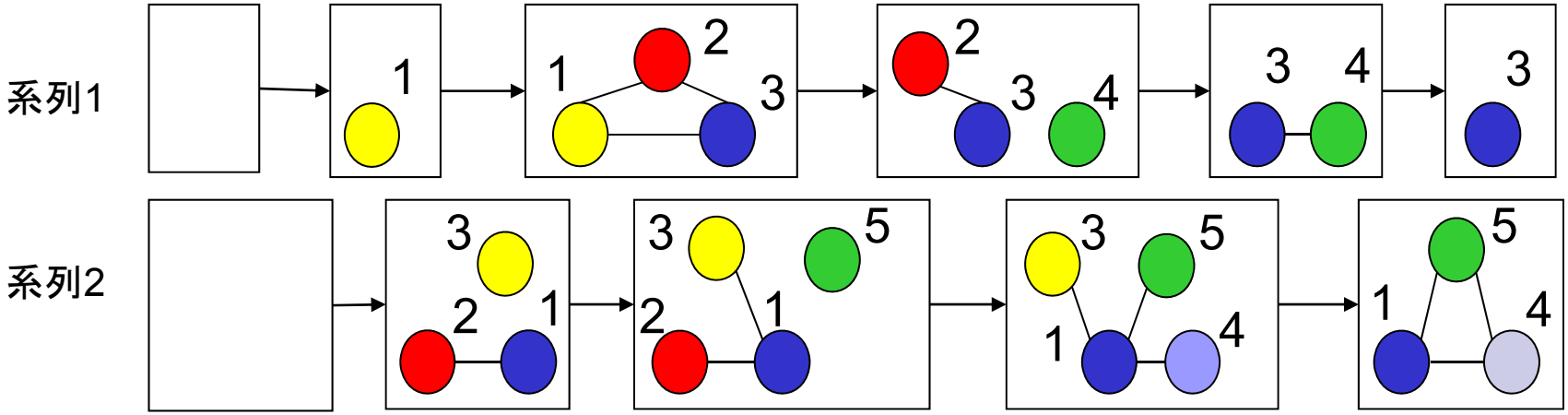
- グラフ系列集合 $DB = \{d_i \mid d_i = \langle g_i^{(1)} g_i^{(2)} \cdots g_i^{(n_i)} \rangle\}$と閾値（最小支持度）σ'が入力として与えられたとき，DB中の頻出するグラフ系列パターンを全て列挙すること



GTRACE [ICDM 2008] の基本アイデア

仮定

グラフ系列中の連続する2つのグラフの間では，構造が大きく変化することはなく，ごく一部の構造のみが変化する．



コンパイル

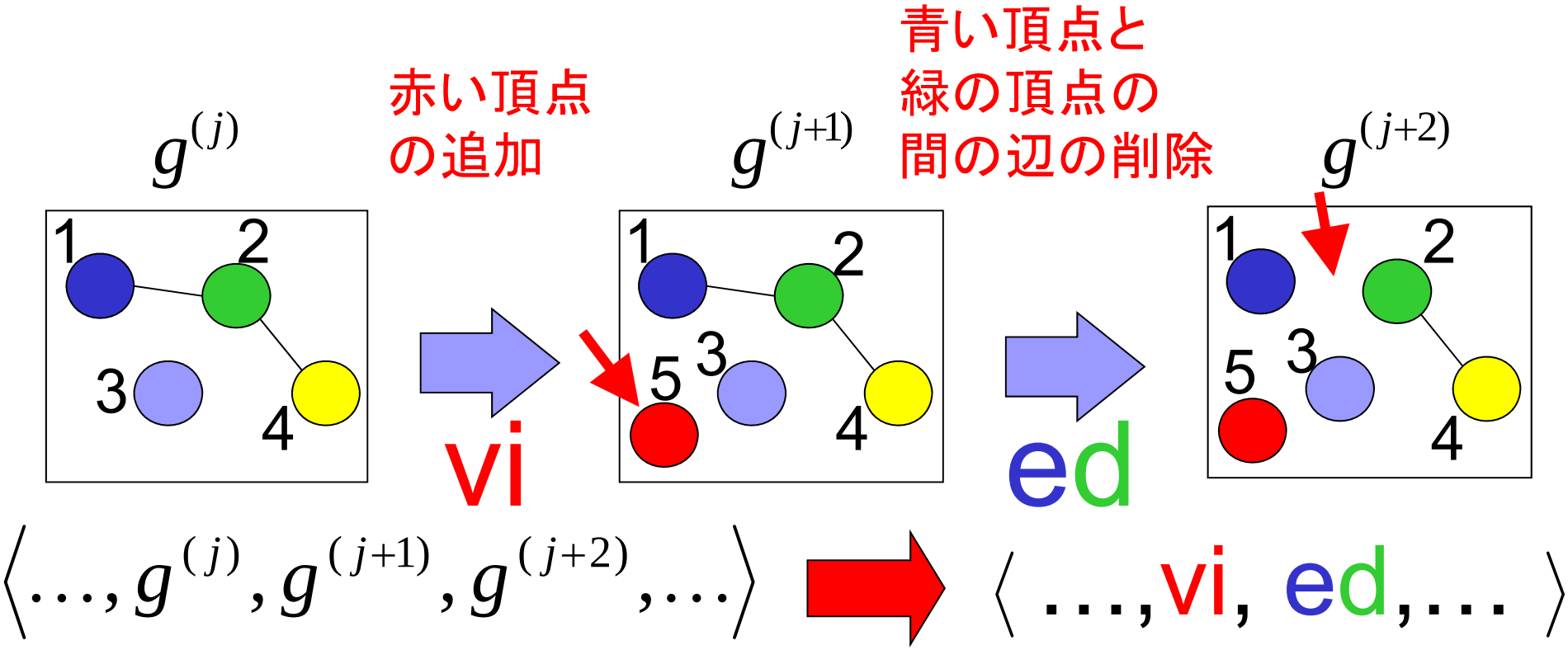
〈(vi),(vi,vi,ei,ei,ei),(vi,ed,ed,vd),(ei,ed,vd),(ed,vd)〉
〈(vi,vi,vi,ei),(vi,ei),(vi,ei,ei,ed,vd),(ei,ed,vd)〉

系列パターンマイニング

頻出部分系列 (FTS) 〈(vi,vi,ei),vi,(ei,ed,vd)〉

グラフの変換

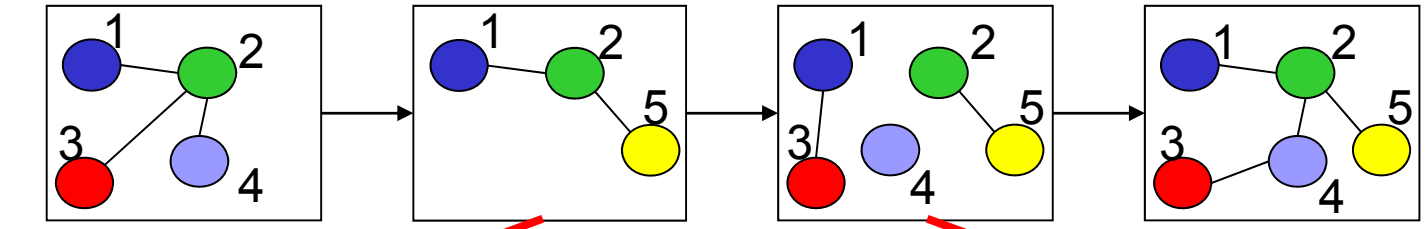
- 頂点や辺の追加，削除，ラベル変更をグラフの変化．



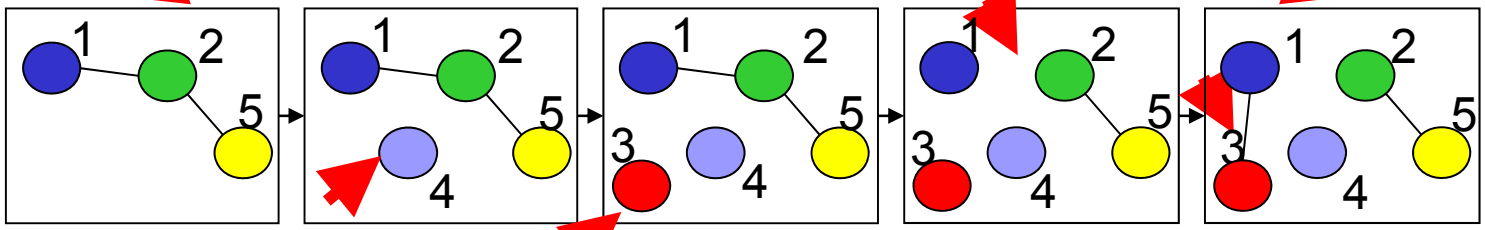
- 6種の変換規則
  - 頂点追加 (VI)，頂点削除 (VD)，頂点ラベル変更(VR)
  - 辺追加(EI)，辺削除(ED)，辺ラベル変更(ER)

グラフ系列の補間

観測されたグラフ系列



補完系列



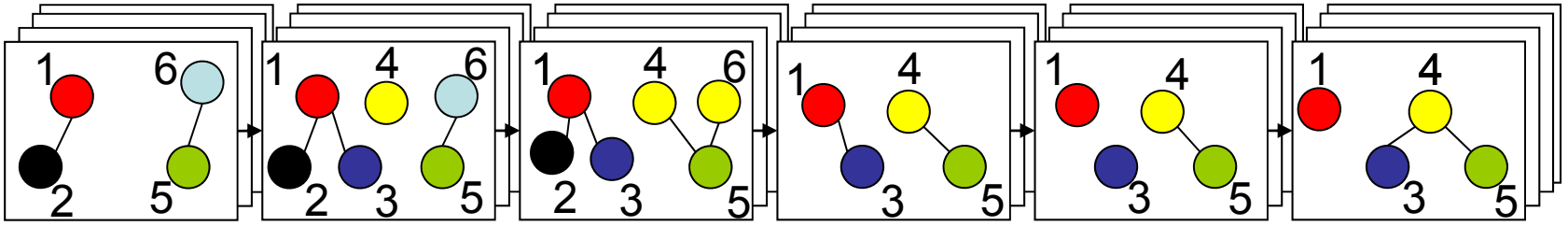
〈…,(vi, vi, ed, ei),…〉

グラフの変化をアイテム集合（変換規則の集合）の系列に変換後，系列パターンマイニングアルゴリズムを適用し，FTSを列挙する．

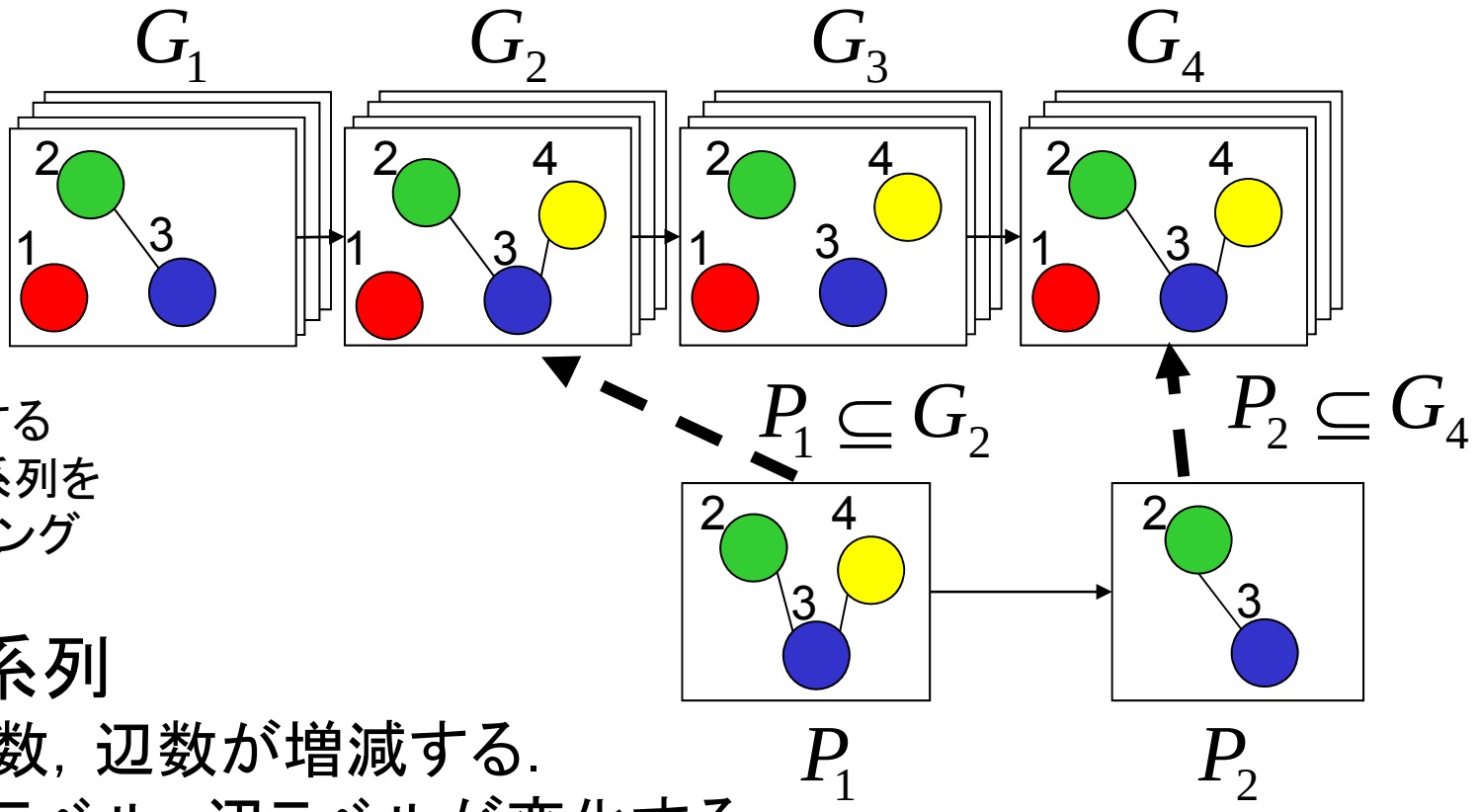
GTRACEの課題

- GTRACEは観測されたグラフ系列中の連続する2つのグラフで，その大部分は変化せず，ごく一部の構造が変化することを仮定

- 観測されたグラフ系列中の連続する2つのグラフが大きく変化する場合には，変換規則の系列が長くなり，膨大な計算時間を要する.



グラフ系列のマイニング

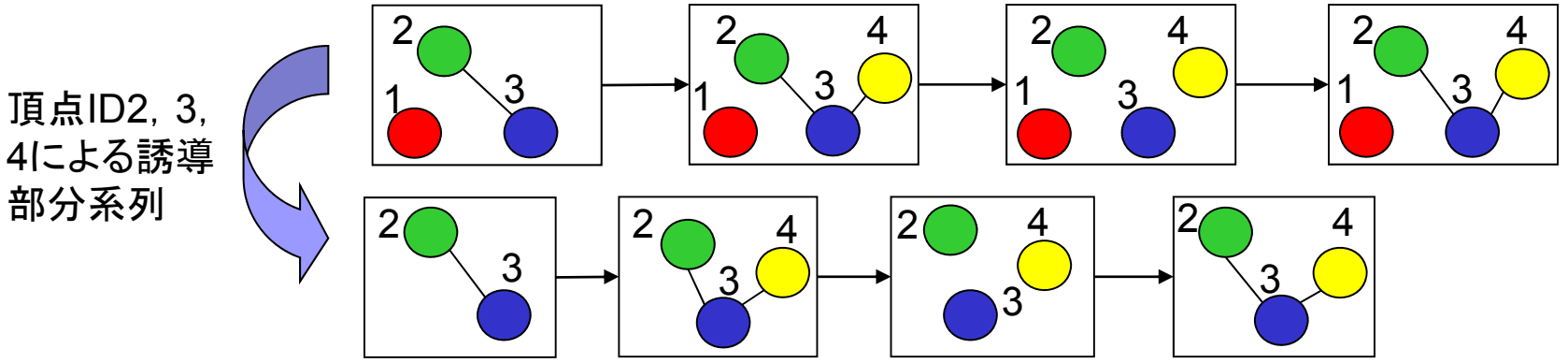


頻出する
部分系列を
マイニング

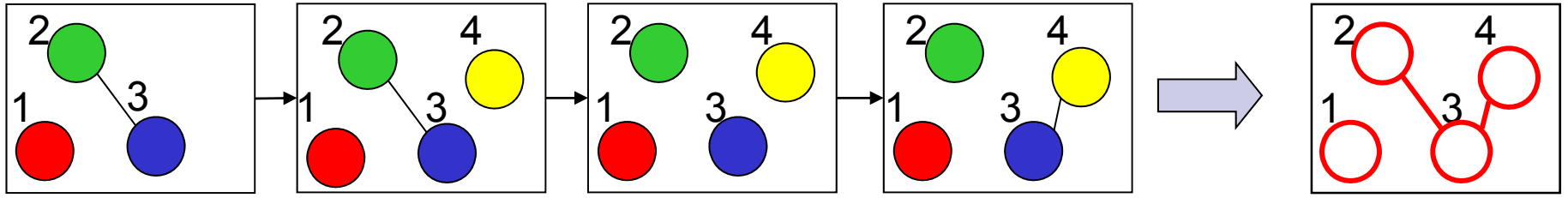
- グラフ系列
  - 頂点数，辺数が増減する．
  - 頂点ラベル，辺ラベルが変化する．
  - 各頂点は，IDをもつ．
  - グラフ系列中の連続する2つのグラフの間で，**構造が大きく変化する．**
  - 個々のグラフが大きく，系列が長い．

頻出関連誘導部分グラフ系列

- FRISS (Frequent, Relevant, and Induced Subgraph Subsequence)
  - 誘導部分グラフ系列

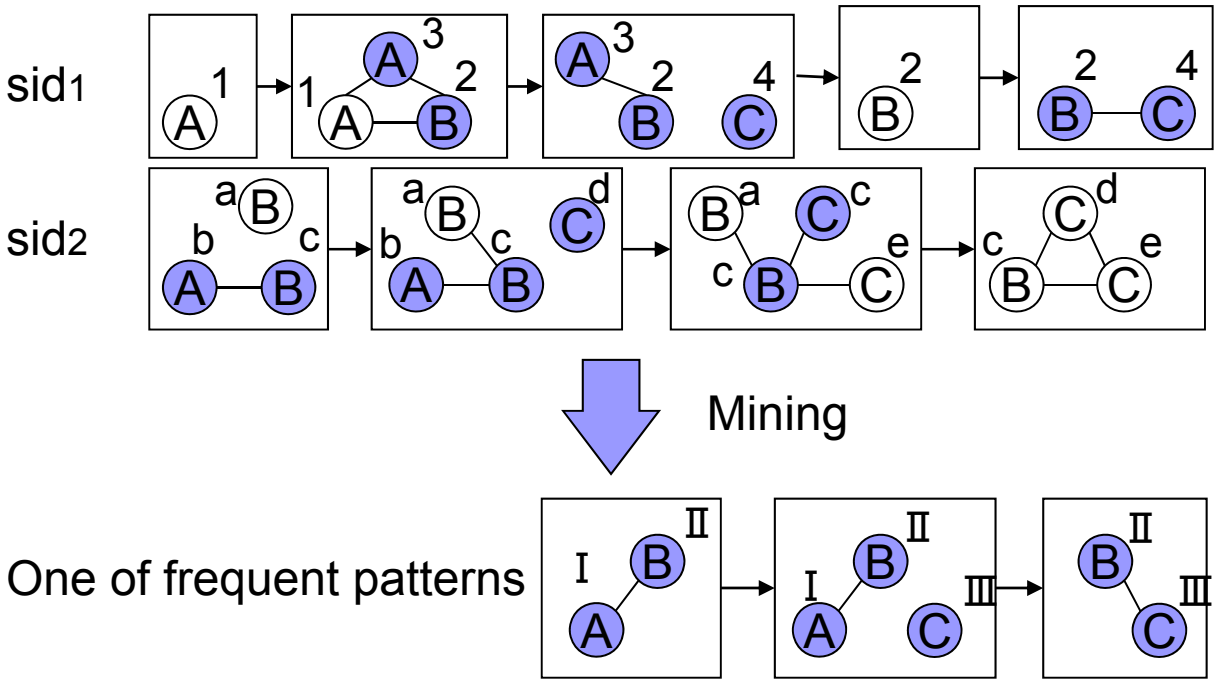


  - 関連部分グラフ系列
    - グラフ系列の和グラフが連結であるグラフ系列

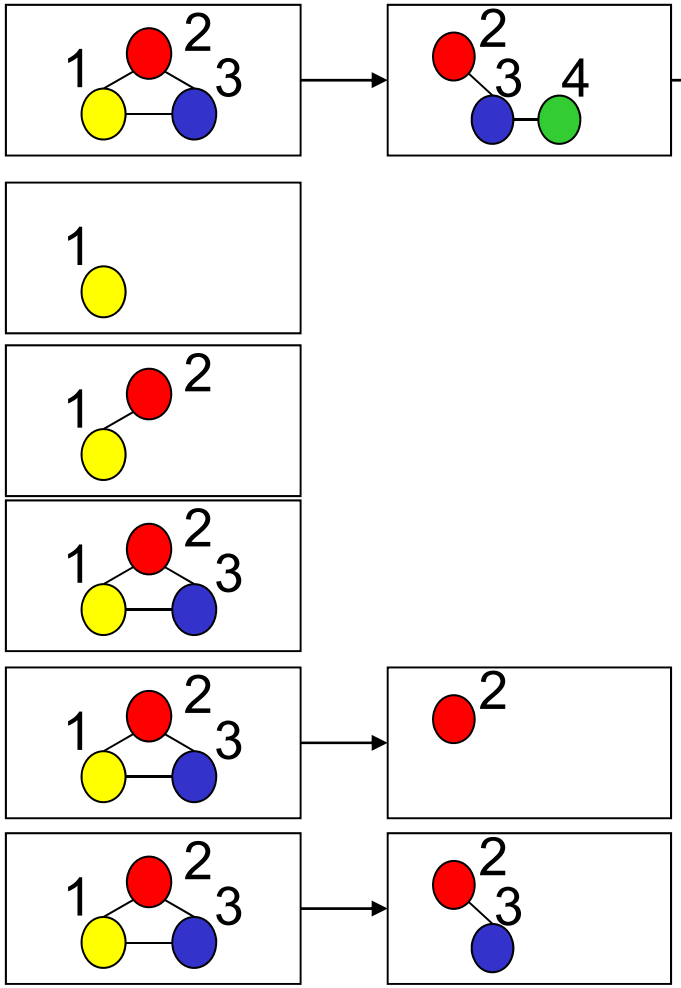


FRISSMinerの問題定義

- グラフ系列集合 $DB = \{d_i \mid d_i = \langle g_i^{(1)} g_i^{(2)} \cdots g_i^{(n_i)} \rangle\}$と閾値（最小支持度）σ'が入力として与えられたとき，DB中の頻出するグラフ系列パターンを全て列挙すること
  - ただし，
    - パターンは元のグラフに誘導部分グラフ系列として含まれているものとする．
    - パターンの和グラフは連結であるとする．



FRISSをマイニングする素朴な探索方法

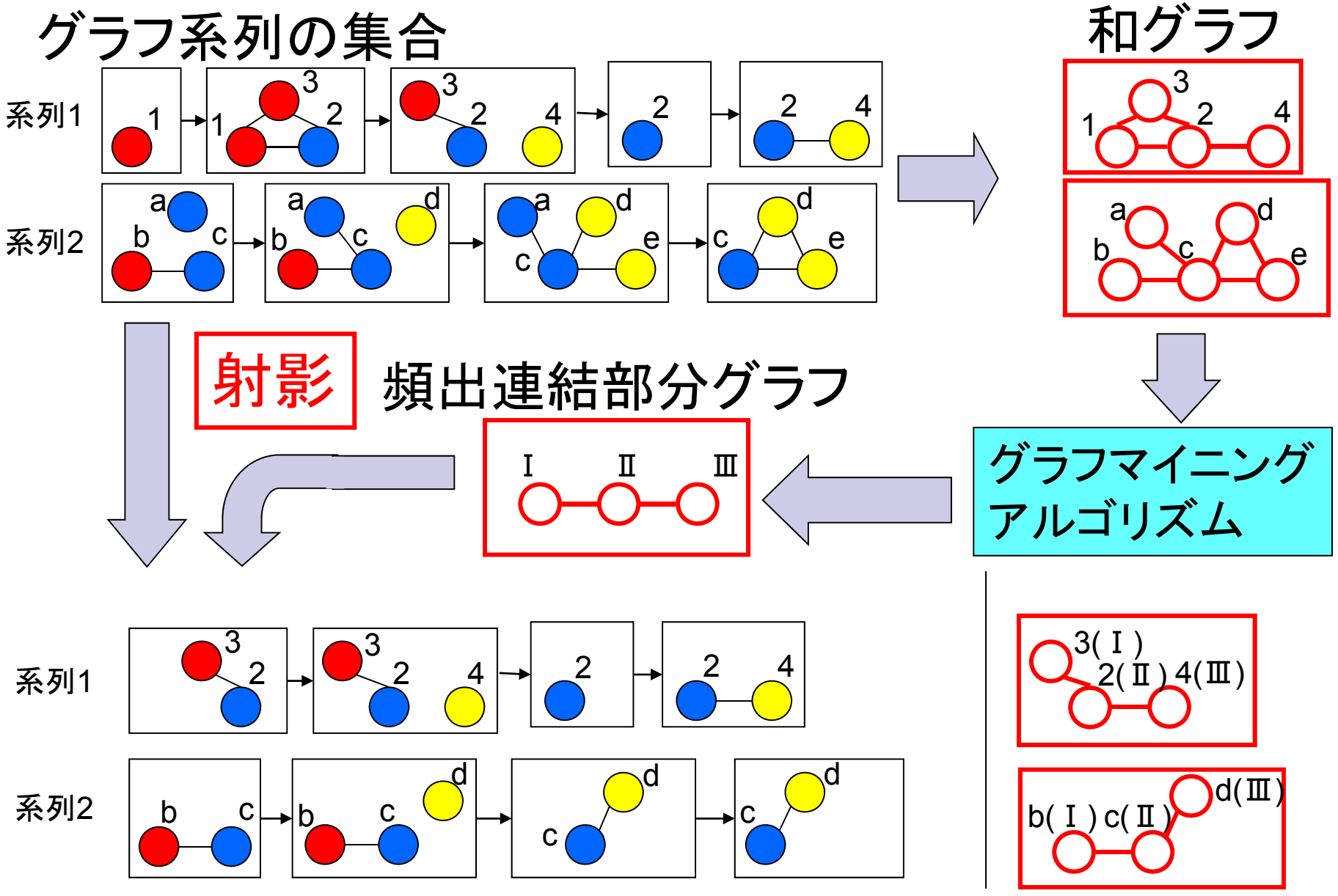


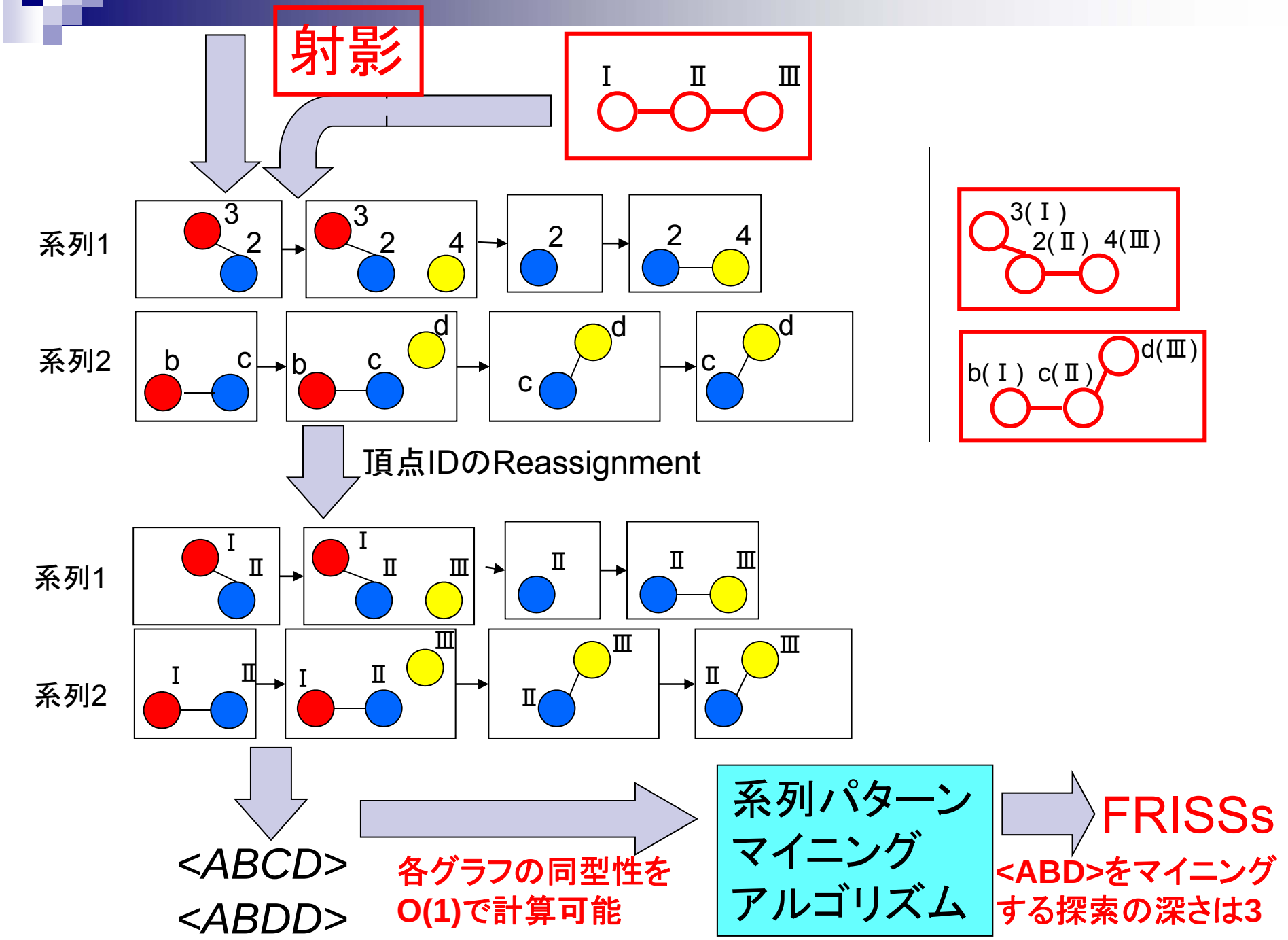
- 頂点を1つずつ追加し、パターンの候補を生成し，頻度を計算する.
- 再帰的に頂点を追加し，頻度が閾値（最小支持度）を下回る場合，バックトラック

bは探索木の分岐数

- 関連部分グラフ系列でないパターンがマイニングされる
- 探索木の深さdが深くなる（深さ優先探索のメモリ使用量　O(bd)）

FRISSMinerのマイニング手順

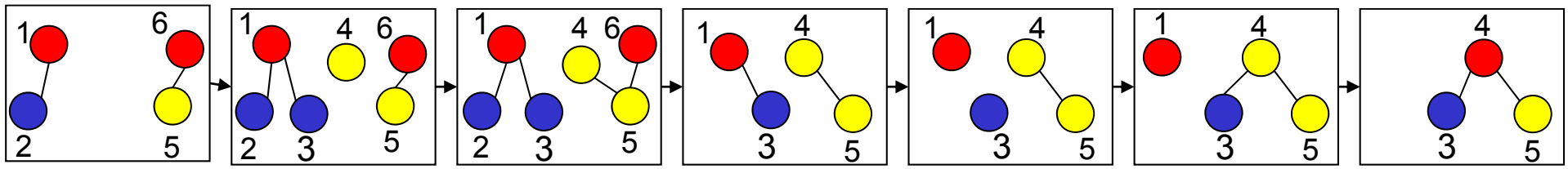



射影
Ⅰ Ⅱ Ⅲ
系列1
系列2
3 2
3 2 4
2
2 4
b c
b c d
c d
c d
3(Ⅰ)
2(Ⅱ) 4(Ⅲ)
b(Ⅰ) c(Ⅱ) d(Ⅲ)
頂点IDのReassignment
系列1
系列2
Ⅰ Ⅱ
Ⅰ Ⅱ Ⅲ
Ⅱ
Ⅱ Ⅲ
Ⅰ Ⅱ
Ⅰ Ⅱ Ⅲ
Ⅱ Ⅲ
Ⅱ Ⅲ
<ABCD>
<ABDD>
各グラフの同型性を
O(1)で計算可能
系列パターン
マイニング
アルゴリズム
FRISSs
<ABD>をマイニング
する探索の深さは3

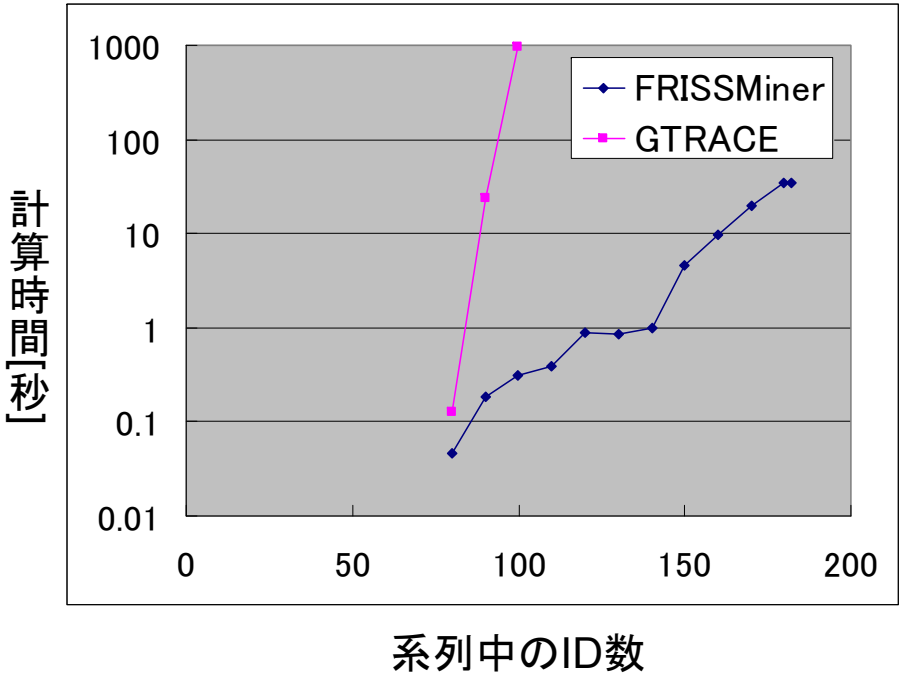
実験結果（実世界データ）

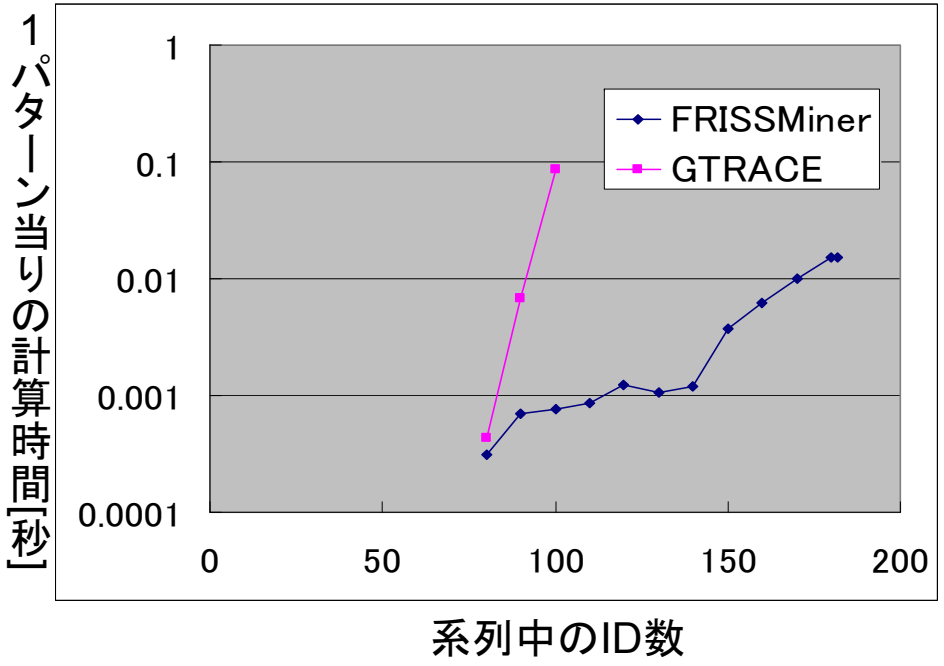
- エンロンデータ
  - 電子メールの履歴データ
  - 123週
- 前処理
  - 頂点ID: E-mailアドレス (人)
  - 辺: ある日にコミュニケーションをとった2名
  - 頂点ラベル: 8種のラベルのいずれか
    - CEO, Employee, Director, Manager, Lawyer, President, Trader, Vice President

グラフ系列データ



実験結果





- グラフ系列中のID数の増加に対して，計算時間は指数関数的に増加（右図）
- グラフ系列中のID数の増加に対して，1パターン当りの計算時間は指数関数的に増加（左図）
- 提案手法FRISSMinerは，グラフ系列が長く，グラフ系列の各グラフが大きいグラフにも適用可能

まとめ

- マイニングするパターンを頻出関連誘導部分グラフ系列(FRISS)に制限することによる効果
  - FRISSは理解困難ではない.
    - FRISSを理解するためにデータベース中のグラフ系列をみる必要がない.
  - 誘導部分グラフ系列に限定することで，パターンの候補が減るため，計算時間は短くなる.
  - 関連部分グラフ系列に限定することで，パターンの候補が減るため，計算時間は短くなる.
  - 射影後のグラフ系列をアイテムの系列で，表現できるので，効率良くパターンをマイニングできる.