

グラフ・部分グラフ同型判定の原理

2011年11月18日

北海道大学大学院 情報科学研究科
 湊ERATO連携講座大学院特論科目(1)
 集中講義「大規模離散計算科学特論」
 客員教授 鷲尾隆(大阪大学産業科学研究所)

1

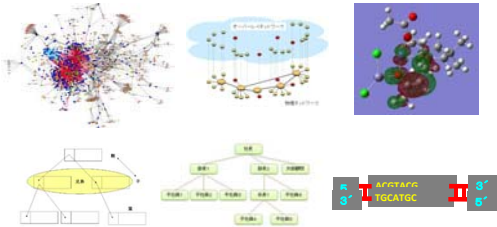
講義内容

- データマイニングとは? — 概論 —
- **グラフ同型・部分グラフ判定の原理**
- 頻出部分グラフマイニングの原理と応用
- グラフ検索の原理と応用
- 統計的因果推論の原理と応用

2

背景

- 世の中にネットワークや木、系列など構造を持つデータが沢山蓄積されている。



- 蓄積されたデータから特定の構造を持つものを検索したり、特徴的部分構造を見つけたい。
- 多くの構造はグラフで表すことができるので、ここではグラフを扱う。

3

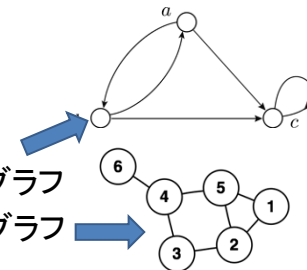
グラフとは?

グラフ $G(V,E)$

- 頂点集合 V
- 辺集合 $E \subseteq V \times V$

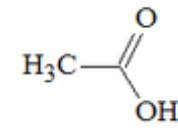
辺に向きがあるとき有向グラフ

辺に向きがないとき無向グラフ



レベル付きグラフ $G(V,E,L,\ell)$

- ラベリング関数 $\ell: \{V,E\} \rightarrow L$
- 右の例 $L = \{H, C, O, -, =\}$



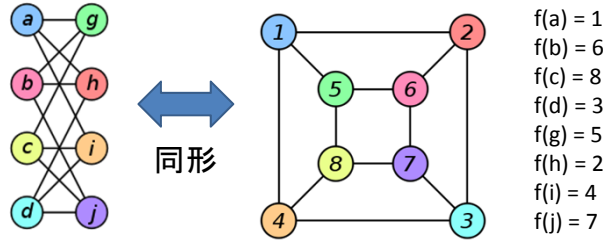
グラフ同型

2つのグラフ $G_1(V_1, E_1, \ell_1, L_1), G_2(V_2, E_2, \ell_2, L_2)$

全単射 $f: V_1 \leftrightarrow V_2$

s.t. $\{v_a, v_b\} \in E_1 \Leftrightarrow \{f(v_a), f(v_b)\} \in E_2, \ell_1(v) = \ell_2(f(v))$

が存在するとき、 G_1 と G_2 は同形であるという。

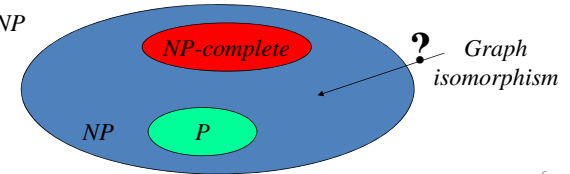


5

グラフ同型判定問題

- 2つのグラフ $G_1(V_1, E_1), G_2(V_2, E_2)$ が与えられたとき、これらが同形か否かを判定する問題。
- 問題の計算複雑性: NP問題に含まれるが、もし $P \neq NP$ ならば P と NP完全のどちらに属するかは未解明。

If $P \neq NP$



6

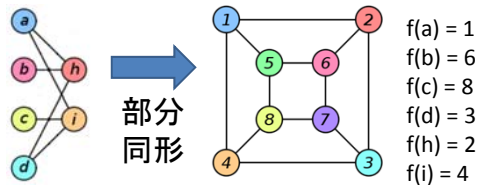
部分グラフ同型

2つのグラフ $G_1(V_1, E_1, \ell_1, L_1), G_2(V_2, E_2, \ell_2, L_2)$

単射 $f: V_1 \rightarrow V_2$

s.t. $\{v_a, v_b\} \in E_1 \Rightarrow \{f(v_a), f(v_b)\} \in E_2, \ell_1(v) = \ell_2(f(v))$

が存在するとき、 G_1 は G_2 と部分グラフ同型である。

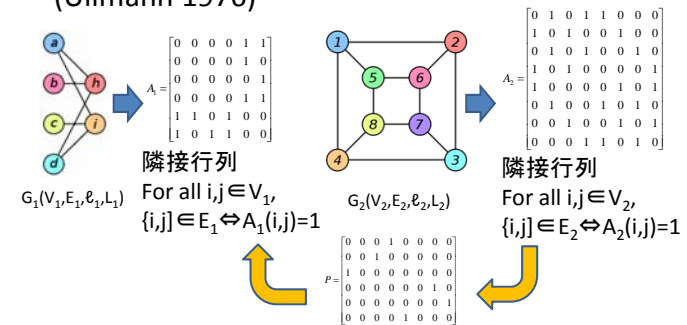


部分グラフ同型は、NP-completeであることが分かっている。

7

(部分)グラフ同型判定アルゴリズム

- (Ullmann 1976)

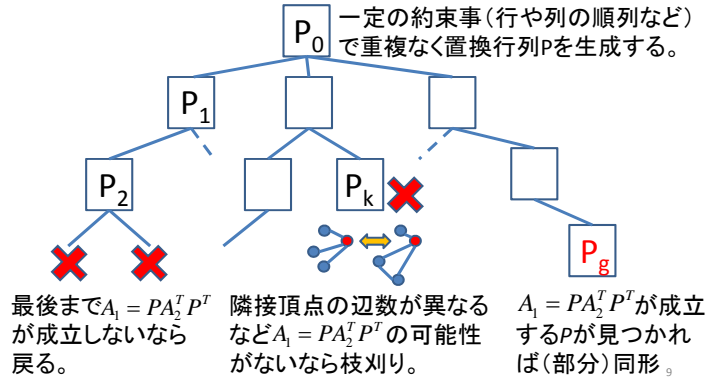


$A_1 = PA_2^T P^T$ となる置換行列 P を探す。
 P が存在すれば(部分)グラフ同形。

8

(部分)グラフ同型判定アルゴリズム

- (Ullmann 1976) (つづき)



グラフ同型判定アルゴリズム

- SD (Schmidt & Druffel 1976)

有向グラフの場合

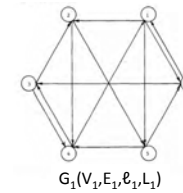


TABLE II. DISTANCE MATRIX FOR GRAPH 1

	1	2	3	4	5	6
1	0	1	2	2	1	1
2	2	0	2	1	1	2
3	2	1	0	1	2	3
4	1	2	1	0	2	2
5	2	3	2	1	0	1
6	1	2	1	2	2	0

(i,j)要素は頂点iと頂点jの最短エッジ距離

TABLE III. ROW CHARACTERISTIC MATRIX FOR GRAPH 1

	1	2	3	4	5
1	3	2	0	0	0
2	2	3	0	0	0
3	2	2	1	0	0
4	2	3	0	0	0
5	2	2	1	0	0
6	2	3	0	0	0

RCM (i,j)要素は頂点iからエッジ距離にある頂点数

TABLE IV. COLUMN CHARACTERISTIC MATRIX FOR GRAPH 1

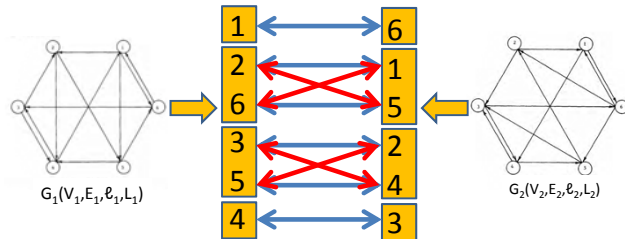
	1	2	3	4	5
1	2	3	0	0	0
2	2	2	1	0	0
3	2	3	0	0	0
4	3	2	0	0	0
5	2	3	0	0	0
6	2	2	1	0	0

CCM (i,j)要素は頂点iまでエッジ距離にある頂点数

グラフ同型判定アルゴリズム

- SD (Schmidt & Druffel 1976) (つづき)

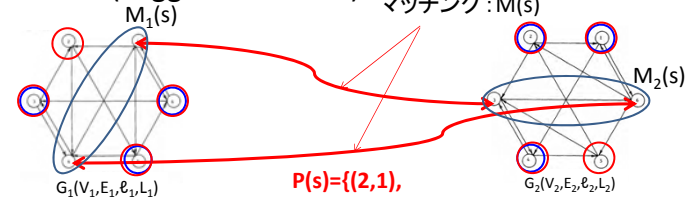
前記RCM,CCM両方の行が完全に同じで区別できない頂点同士にグループ化。



区別できない頂点同士のみ照合してグラフが同形か否かを調べる。照合回数は6!=720回から2×2=4回に削減。

(部分)グラフ同型判定アルゴリズム

- VF2 (Foggia et al. 2001) 探索ステップsのマッチング: M(s)



$P(s) = \{(2,1), (3,1), (5,1), (6,1)\}$

$T_1^{out}(s)$: M(s)内の頂点から指されている頂点
 $T_1^{in}(s)$: M(s)内の頂点を指している頂点
 $T_2^{out}(s)$: M(s)内の頂点から指されている頂点
 $T_2^{in}(s)$: M(s)内の頂点を指している頂点

次のマッチング候補 $P(s) = T_1^{out}(s) \times \min T_2^{out}(s)$ or $T_1^{in}(s) \times \min T_2^{in}(s)$ or $(V_1 - M_1(s)) \times \min(V_2 - M_2(s))$ (minは頂点ラベルなど辞書順で最小)

(部分) グラフ同型判定アルゴリズム

- VF2 (Foggia et al. 2001) (つづき)

$F(s, (i, j))$: $P(s)$ から更に可能性のある対応 (i, j) を選ぶ関数

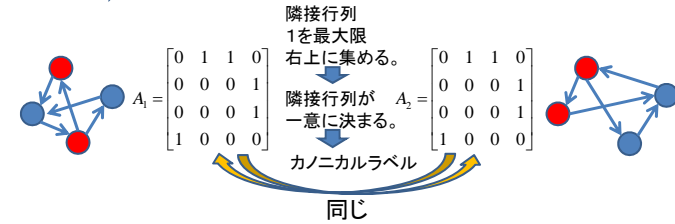
- $(i, j) \in P(s)$ の i と $M_1(s)$ 内頂点との辺, j と $M_2(s)$ 内頂点との辺が一致するかを確認。
- $(i, j) \in P(s)$ の i のラベルと j のラベルが一致するかを確認。
- $T_1^{out}(s)$ と $T_2^{out}(s)$, $T_1^{in}(s)$ と $T_2^{in}(s)$, $V_1 - M_1(s)$ と $V_2 - M_2(s)$ に含まれる頂点数が一致する(或いは以下)か確認。

選ばれた対応 (i, j) を $M(s)$ に含める。
 新たな $M(s)$ で(部分)同形対応が取れるかを確認。
 対応が取れないなら $P(s)$ と $F(s, (i, j))$ の計算を繰り返す。

13

グラフ同型判定アルゴリズム

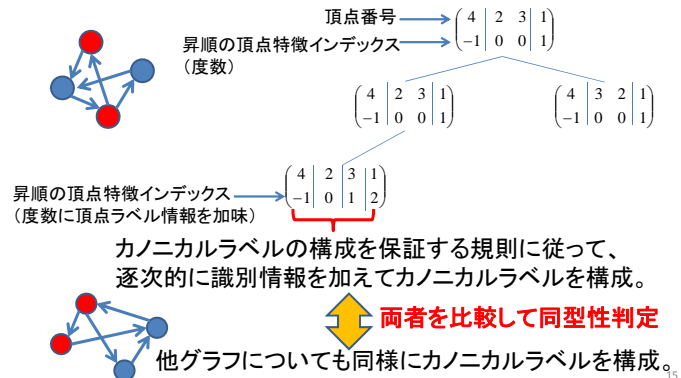
- Naughty (McKay 1981)
 - グラフ同士を直接比較するのではなく、各グラフの**カノニカルラベル**を比較する。
 - **カノニカルラベル**: 1つのグラフに一意に対応する。
- ➡ **カノニカルラベルが同じならグラフは同型**



ただし、行列の置換変換でカノニカルラベルを探すのは計算が大変。

グラフ同型判定アルゴリズム

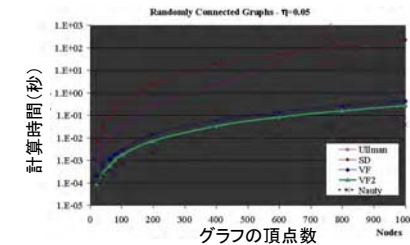
- Naughty (McKay 1981) (つづき)



15

(部分) グラフ同型判定アルゴリズムの比較

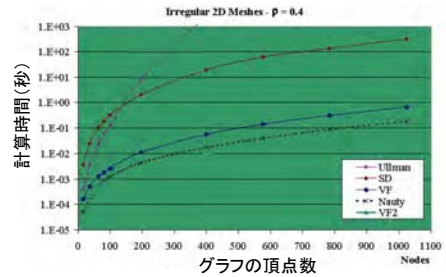
- Ullmann, SD, VF, VF2, Naughty の5つのアルゴリズムを人工データで比較 (Foggia et al. 2001)
 - 実装: C++, Intel Celeron 500 Mhz PC, 128 MB RAM
 - データ: ランダム連結グラフ, 辺存在確率0.05
- 1000ペアグラフの平均判定時間を計算



16

(部分)グラフ同型判定アルゴリズムの比較

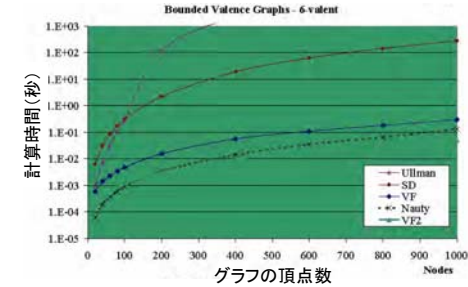
- データ: 非正規平面グラフ, 非正規辺存在確率0.4
1000ペアグラフの平均判定時間を計算



17

(部分)グラフ同型判定アルゴリズムの比較

- データ: Bounded Valenceグラフ, 頂点度数6
1000ペアグラフの平均判定時間を計算



18

グラフ・部分グラフ同型判定: まとめ

- 以下の2つの原理が用いられる。
 - グラフの各部分の特徴づける量を比較しながら、マッチング探索空間を絞り込み同形性判定を行う方法
 - 各グラフと一意に対応するカノニカルラベルを計算し、それらを比較して同型性判定を行う方法
- どちら方法も同程度の効率を持つ。
- 現状ではVF2(前者の方法)とNauty(後者の方法)の効率が拮抗している。

19