Vol.2017-DBS-164 No.11
2017/1/17

# 辺色付きグラフのクラスタリングにおけるクラスタ表現決定法

薄永　光平[1]　上土井　陽子[2]　若林　真一[2]

**概要**：頂点間の二項関係が categorical(種別) である辺ラベル付きグラフの新しいクラスタリング問題がある [1]。クラスタリングとは、頂点の集合をクラスタという部分集合に分割することである。クラスタリングの目標は、オブジェクトの集合を同じクラスタ内のオブジェクトが、他のクラスタ内のオブジェクトより互いに似ているように異なるクラスタに分割することである。辺色付きグラフのクラスタリングでは、クラスタ内のラベルができるだけ共通している辺を多くクラスタリングし、かつ存在しない辺がなるべくクラスタ間になるようにクラスタリングすることを目的としている。応用分野として、ソーシャルネットワーク、タンパク質間相互作用ネットワーク、文献データベースのクラスタリングによるトピックに注目した代表的なグループ発見が挙げられる。本稿では、辺色付きグラフのクラスタリングにおいて、各クラスタを表現するラベルの集合を決定するクラスタラベル決定アルゴリズムを提案し、その最適性を証明する。

## 1. はじめに

従来では、グラフの頂点間の関係を実数値で表現していた。本研究では、頂点間の関係を色（ラベル）で表現した問題を扱う。頂点間の関係を色 (ラベル) で表したグラフを辺ラベル付きグラフと呼ぶ。辺ラベル付きグラフの研究は、多くの実際のアプリケーションによって動機づけされ、データマイニング文学において関心が高まりつつある [2,3,4]。例として、ソーシャルネットワークは、一般に頂点が個人を表すグラフで、これらの個人間の辺モデル関係として表される。

図 1 にソーシャルネットワークでの個人間の関連を辺ラベル付きグラフとして表現した例を示す。ここで、頂点は個人を表し、辺は個人間の関係、例えば、共通の話題を表現しているとする。さらに、$(x1, x3)$, $(x1, x5)$, $(x3, x5)$ の 3 つの辺がスポーツ、$(x2, x4)$, $(x2, x6)$, $(x4, x6)$ の 3 つの辺が教育、$(x2, x3)$, $(x3, x4)$, $(x3, x6)$ の 3 つの辺が娯楽、$(x1, x4)$, $(x4, x5)$ の 2 つの辺が仕事を示している。図のようなつながりがグラフとして表されていて、コミュニティでクラスタリングしたいとすると、代表的なものスポーツと教育で分割するとほかの話題を含むことを少なく同じ色で分けられる。クラスタリングができると全体的なグループ分けをしながら、代表的なトピックでそれぞれのグループを表せる。

上記の問題に対して、文献 [1] では辺ラベル付きグラフの詳細な問題の定式化とランダム近似アルゴリズムを提案している。本稿では、辺レベル付きグラフの頂点集合の分割が与えられたときに、クラスタラベルと呼ばれる各クラスタを表現するラベルを割り当てる問題に着目し、最適化アルゴリズムを提案する。

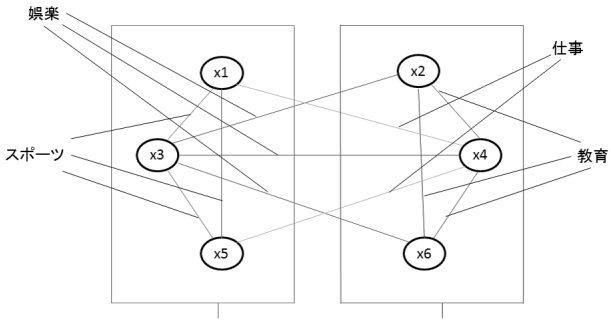


**図 1**　ソーシャルネットワークでの個人間の話題をモデル化した辺ラベル付きグラフの例

## 2. 辺ラベル付きクラスタリング問題

本節ではまず、本研究で対象とする辺ラベル付きグラフについて紹介する。辺ラベル付きグラフには、辺が一種類のラベルをもつ場合と、多種類のラベルをもつ場合がある。次に文献 [1] で定義されている 2 つのクラスタリング問題に

[1] 広島市立大学情報科学部,
〒 731-3194 広島市安佐南区大塚東三丁目 4-1
[2] 広島市立大学大学院情報科学研究科,
〒 731-3194 広島市安佐南区大塚東三丁目 4-1

ついて紹介する。1つ目は辺ラベルが1種類の場合の辺ラベル付きグラフのクラスタリング問題である CHROMATIC-CORRELATION CLUSTERING 問題である。2つ目は辺ラベルが多種類の場合の辺ラベル付きグラフのクラスタリング問題である MULTICHROMATIC-CORRELATION CLUSTERING 問題である。

### 2.1 CHROMATIC-CORRELATION-CLUSTERING 問題

次に、頂点間の関係が色で表現されている場合のグラフの分割問題である CHROMATIC-CORRELATION-CLUSTERING を考える。入力は、辺ラベル付きグラフ $G = (V, E, L, l_0, \ell)$。ここで、$V$ は、頂点の集合で、$V_2$ の部分集合 $E$ は、辺の集合で $L$ は、ラベルの集合で $l_0$ は、$(x, y) \notin E$ の場合にのみ、$\ell(x, y) = \{l_0\}$ となる。$\ell : V_2 \rightarrow L \cup \{l_0\}$ は、$V$ 内の各二項組にラベルを割り当てる関数である。出力は、目的関数を最小化するクラスタリング $C : V \rightarrow N$ とクラスタラベル関数 $c\ell : C[V] \rightarrow L$ を見つけることである。目的関数は、以下の式 (1) である。

$$cost(G, C, cl) = \sum_{(x,y)\in V2, C(x)=C(y)} (1 - I[\ell(x, y) = c\ell(C(x))]) + \sum_{(x,y)\in V2, C(x)\neq C(y)} (1 - I[\ell(x, y) = l_0]) \quad (1)$$

式 (1) は、二つの項で構成されており、それぞれクラスタ内とクラスタ間のコストを表している。$C$ は各頂点にグループの番号を割り当てる関数で、$c\ell : C[V] \rightarrow L$ は、クラスタにラベルを与える関数である。また、$I[\cdot]$ は、指示関数を示しており、$I$ の引数は命題であり、命題が真なら 1、偽なら 0 を返す。

CHROMATIC-CORRELATION-CLUSTERING 問題が一般的なグラフのクラスタリング問題と大きく異なるところは、クラスタにラベルを付ける関数 $c\ell : C[V] \rightarrow L$ があることである。CHROMATIC-CORRELATION-CLUSTERING 問題は、一般的なグラフのクラスタリング問題と同様に、$NP$ 困難である [1]。

### 2.2 MULTICHROMATIC-CORRELATION-CLUSTERING 問題

次に、頂点間の関係が複数のラベルで表すことができる場合のグラフのクラスタリング問題である MULTICHROMATIC-CORRELATION-CLUSTERING 問題を考える。この問題の入力は、辺が複数のラベルをもっている辺ラベル付きグラフである。以下にこの問題の定義を示す。入力は、辺ラベル付きグラフ $G = (V, E, L, l_0, \ell)$ である。ここで、$V$ は、頂点の集合で、$V_2$ は、全ての頂点の二項組の組み合わせである。$V_2$ の部分集合 $E$ は、辺の集合で $L$ は、ラベルの集合で $l_0$ は、$(x, y) \notin E$ の場合にのみ、$\ell(x, y) = \{l_0\}$ となる。$\ell : V_2 \rightarrow 2^L \cup \{l_0\}$ は、$V$ 内の各二項組にラベルを割り当てる関数である。$d_\ell : 2^L \cup \{l_0\} \times 2^L \cup \{l_0\} \rightarrow R^+$ は、ラベルの集合間の距離を正の実数で出力する関数である。出力は、目的関数を最小化するクラスタリング $C : V \rightarrow N$ とクラスタラベル関数 $c\ell : C[V] \rightarrow 2^L$ を見つけることである。目的関数は、以下の式 (2) である。ここで $2^L$ は、$L$ のべき集合から空集合を除いた集合で $2^L = \{S \subseteq L \mid |S| \geq 1\}$ である。

$$cost(G, C, c\ell) = \sum_{(x,y)\in V_2, C(x)=C(y)} d_\ell(\ell(x, y), c\ell(C(x))) + \sum_{(x,y)\in V_2, C(x)\neq C(y)} d_\ell(\ell(x, y), \{l_0\}) \quad (2)$$

$C$ は、各頂点にグループの番号を割り当てる関数で、$c\ell : C[V] \rightarrow 2^L$ は、頂点に割り当てたグループ番号にラベルとして $L$ の部分集合を割り当てる関数である。

MULTICHROMATIC-CORRELATION-CLUSTERING 問題が、CHROMATIC-CORRELATION-CLUSTERING 問題の一般化であることが簡単にわかる。入力グラフの全ての辺が単一ラベルのとき、MULTICHROMATIC-CORRELATION-CLUSTERING 問題は、CHROMATIC-CORRELATION-CLUSTERING 問題に帰着する。そのとき、距離 $d_\ell$ は、以下のように定義される。

$$d_\ell(\{l_1\}, \{l_2\}) = \begin{cases} 0 \; if \; l_1 = l_2 \\ 1 \; otherwise \end{cases} \quad (3)$$

複数ラベルと単一ラベルの定式化の間の差別化のキーポイントは、前者はラベル集合は互いにどのくらい違うのかを測るために距離関数 $d_\ell$ を利用することである。$d_\ell$ は様々な選択肢の中で本研究では、簡単かつ有効性の間の良いトレードオフとして一般的なハミング距離を使用する。ラベルの 2 つの集合の間のハミング距離 $L_1 \subseteq L$, $L_2 \subseteq L$ は、$L_1$ と $L_2$ の間の不一致の数として以下のように定義される。

$$d_\ell(L_1, L_2) = |L_1 - L_2| + |L_2 - L_1| \quad (4)$$

[MULTICHROMATIC-CORRELATION-CLUSTERING 問題の例と許容解]

頂点の集合 $V$ は、$V = \{x1, x2, x3, x4, x5, x6\}$、辺の集合 $E$ は、$E$=$\{\{x1, x3\}, \{x1, x4\}, \{x1, x5\}, \{x2, x3\}, \{x2, x4\}, \{x2, x6\}, \{x3, x4\}, \{x3, x5\}, \{x3, x6\}, \{x4, x5\}, \{x4, x6\}\}$、ラベルの集合 $L$ は、$L$=$\{g, p, b, r\}$、$V$ 内の各二項組にラベルを割り当てる関数 $\ell$ は、$\ell(x1, x3)$=$\{g\}$, $\ell(x1, x4)$=$\{g, r\}$, $\ell(x1, x5)$=$\{g, p, b\}$, $\ell(x2, x3)$=$\{g, b, r\}$, $\ell(x2, x4)$=$\{g, p\}$, $\ell(x2, x6)$=$\{p\}$, $\ell(x3, x4)$=$\{g, p, b\}$, $\ell(x3, x5)$=$\{g, b, r\}$, $\ell(x3, x6)$=$\{b, r\}$, $\ell(x4, x5)$=$\{b\}$, $\ell(x4, x6)$=$\{p, b\}$ とし、存在しない辺は、点線で表し

ており $\ell(x1,x2)$={$l_0$}, $\ell(x1,x6)$={$l_0$}, $\ell(x2,x5)$={$l_0$}, $\ell(x5,x6)$={$l_0$} とする入力の図 **??** を以下に示す。べき集合 $2^L$ は、$2^L$={{$g$}, {$p$}, {$b$}, {$r$}, {$g,p$}, {$g,b$}, {$g,r$}, {$p,b$}, {$p,r$}, {$b,r$}, {$g,p,b$}, {$g,p,r$}, {$g,b,r$}, {$g,p,b,r$}} となる。クラスタの番号 (名前) をそれぞれ 1, 2 とし、図 2 のようにクラスタリングする場合を考える。1 のクラスタのラベルは {$g,b$} となり、2 のクラスタのラベルは {$p$} となる。全ての辺を目的関数である式 (2) に当てはめると、$d_\ell(\ell(x1,x3), c\ell(C(x1))$=1,$d_\ell(\ell(x1,x4), \{l_0\})$=3,$d_\ell(\ell(x1,x5), c\ell(C(x1))$=1,$d_\ell(\ell(x2,x3), \{l_0\})$=4,$d_\ell(\ell(x2,x4), c\ell(C(x2))$=1,$d_\ell(\ell(x2,x6), c\ell(C(x2))$=0,$d_\ell(\ell(x3,x4), \{l_0\})$=4, $d_\ell(\ell(x3,x5), c\ell(C(x3))$=1, $d_\ell(\ell(x3,x6), \{l_0\})$=3, $d_\ell(\ell(x4,x5), \{l_0\})$=2,$d_\ell(\ell(x4,x6), c\ell(C(x4))$=1 となり、cost の合計は 21 である。

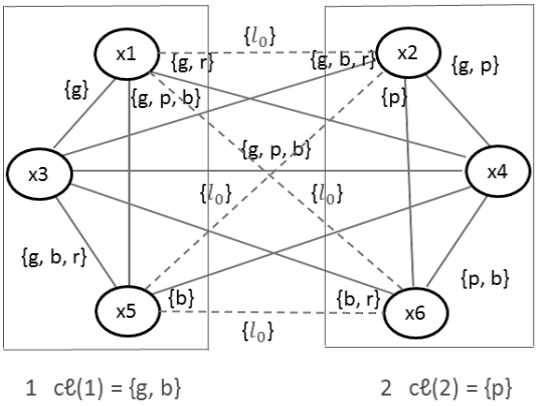


**図 2** MULTICHROMATIC-CORRELATION-CLUSTERING 問題の許容解

上の例からわかることは、出力クラスタリングでは、異なるクラスタ内の頂点の各ペアの理想的な状態は、「辺がない」場合のモデルである $l_0$ のラベルで表される関係をもつことである。また、CHROMATIC-CORRELATION-CLUSTERING 問題の目的は、ほかのラベルを含むことを少なく代表的なラベルでそれぞれのグループを表すことであるのに対し、MULTICHROMATIC-CORRELATION-CLUSTERING 問題の目的は、クラスタ内の辺のラベルができるだけ共通している辺を多く含むようにクラスタリングし、かつ存在しない辺がなるべくクラスタ間になるようにクラスタリングすることである。クラスタラベルは、クラスタを代表するラベルを全て入れる必要があると考えられる。

## 3. 提案クラスタラベル決定法

本節では、多種類の辺ラベルをもつ辺ラベル付きグラフの頂点集合のクラスタリングが与えられたときに、そのクラスタラベルを決定する問題に着目する。そして、与えられた、クラスタリングにおいて、目的関数を最小化するクラスタラベルを決定するアルゴリズムを提案する。また、次節にて提案アルゴリズムの最適性を証明する。

### 3.1 提案クラスタラベル決定アルゴリズム

与えられた、クラスタリングにおいて、目的関数を最小化するクラスタラベルを決定するアルゴリズムである提案決定アルゴリズムの入出力と手順を以下に示す。

入力：辺ラベル付きグラフ $G=(V,E,L,l_0,\ell)$ でのクラスタ $C_i \subseteq V$ とクラスタ $C_i$ による辺ラベル付きグラフ $G$ の誘導部分グラフ $G'=(C_i,E_{C_i},L,l_0,\ell_{C_i}), E_{C_i} \subseteq E$, ただし、$\forall u,v \in C_i[(u,v) \in E_{C_i} \Leftrightarrow (u,v) \in E]$ を満たす。$\ell_{C_i}: E_{C_i} \to 2^L \cup \{l_o\}$, ただし、$\forall e \in E_{C_i}[\ell_{C_i}(e)=\ell(e)]$ を満たす。

出力：目的関数 (式 (2)) を最小化するクラスタラベル $c\ell(C_i) \subseteq L$

提案クラスタラベル決定アルゴリズム

$STEP1$ :$E_{C_i}$ 内の最も多くの辺についているラベル $\ell_{max}$ を 1 つだけを含む集合をクラスタラベル $c\ell(C_i)$ にする ($c\ell(C_i)=\{\ell_{max}\}$)
$\ell_{max}$ となる候補のラベルが複数ある場合は、その中からランダムに 1 つだけ選び、クラスタラベルにする

$STEP2$ :$E_{C_i}$ 内の過半数の辺についているラベルの集合 $M_\ell$ を求めて、$c\ell(C_i) \cup M_\ell$ をクラスタラベルに更新する ($c\ell(C_i) \leftarrow c\ell(C_i) \cup M_\ell$)

### 3.2 提案クラスタラベル決定アルゴリズムを具体例に適用

頂点の集合 $V$ は、$V=\{x1,x2,x3,x4,x5,x6\}$、辺の集合 $E$ は、$E$={{$x1,x3$}, {$x1,x4$}, {$x1,x5$}, {$x2,x3$}, {$x2,x4$}, {$x2,x6$}, {$x3,x4$}, {$x3,x5$}, {$x3,x6$}, {$x4,x5$}, {$x4,x6$}}、ラベルの集合 $L$ は、$L$={{$g$}, {$p$}, {$b$}, {$r$}}、$V$ 内の各二項組にラベルを割り当てる関数 $\ell$ は、$\ell(x1,x3)$={$g$}, $\ell(x1,x4)$={$g,r$}, $\ell(x1,x5)$={$g,p,b$}, $\ell(x2,x3)$={$g,b,r$}, $\ell(x2,x4)$={$g,p$}, $\ell(x2,x6)$={$p$}, $\ell(x3,x4)$={$g,p,b$}, $\ell(x3,x5)$={$g,b,r$}, $\ell(x3,x6)$={$b,r$}, $\ell(x4,x5)$={$b$}, $\ell(x4,x6)$={$p,b$} とし、存在しない辺は、点線で表しており $\ell(x1,x2)$={$l_0$}, $\ell(x1,x6)$={$l_0$}, $\ell(x2,x5)$={$l_0$}, $\ell(x5,x6)$={$l_0$}、そして {$x1,x3,x5$} と {$x2,x4,x6$} に分割し、それぞれの名前を 1,2 とする入力の図 3 がある。

図 3 に提案クラスタラベル決定アルゴリズムを適用する。まず、1 のクラスタに注目すると、$STEP1$ として、最も多い辺のラベル $g$ を一つだけクラスタラベルに $c\ell(1)$ にする ($c\ell(1)=\{g\}$)。$STEP2$ として、過半数のラベルの集合 $M_\ell=\{g,b\}$ を求めて、$c\ell(1) \cup \{g,b\}$ をクラスタラベルに更新する。よって、1 のクラスタラベル $c\ell(1)$ は、{$g,b$} となる。次に、2 のクラスタに注目すると、$STEP1$ として、最も多い辺のラベル $p$ を一つだけクラスタラベル $c\ell(2)$ に

する ($c\ell(2) = \{p\}$)。$STEP2$ として、過半数のラベルの集合 $M_\ell = \{p\}$ を求めて、$c\ell(2) \cup \{p\}$ をクラスタラベルに更新する。よって、2 のクラスタラベル $c\ell(2)$ は、$\{p\}$ となる。

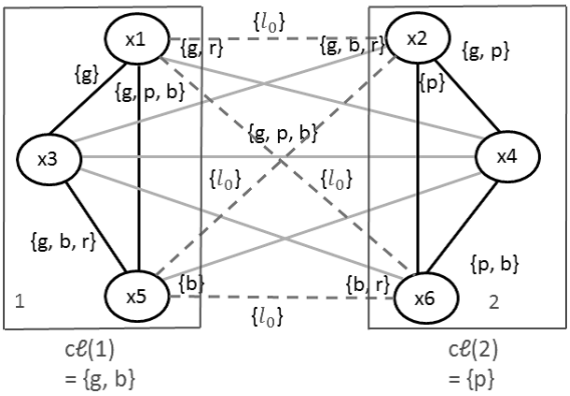


**図 3** 提案クラスタラベル決定アルゴリズムの適用例

## 4. 最適性の証明

提案アルゴリズムの最適性を以下の定理が成り立つことを示すことにより証明する。

[**定理 1**] あるクラスタ $C_i$ において、(1)$C_i$ 内の辺に最も多く付けられているラベル $\ell_{max}$ を 1 つ含み、かつ (2)$C_i$ 内の辺の過半数が共通して持つ全てのラベル $\ell_j$ がクラスタラベル $c\ell(C_i)$ に含まれ、(3)(1) で選ばれた $\ell_{max}$ 以外で、辺の半数未満しか共通していない、どのラベル $\ell_k$ もクラスタラベル $c\ell(C_i)$ に含まれない場合、かつ、その時に限り、そのクラスタ $C_i$ の $cost$ は最小である。

[**証明概略**] 命題を場合分けし、証明するため、各条件を以下のように記号 $P, Q, W, R$ を用いて記す。

$W$ : (1)$C_i$ 内の辺に最も多く付けられているラベル $\ell_{max}$ を 1 つ含む。

$P$ : (2)$C_i$ 内の辺の過半数が共通して持つ全てのラベル $\ell_j$ がクラスタラベル $c\ell(C_i)$ に含まれる。

$Q$ : (3)$W$ で選ばれた ((1) で選ばれた)$\ell_{max}$ 以外で、辺の半数未満しか共通していない、どのラベル $\ell_k$ もクラスタラベル $c\ell(C_i)$ に含まれない。

$R$ : クラスタ $C_i$ の $cost$ は最小である

$(R \Rightarrow P \wedge Q) \Leftrightarrow (\neg R \vee (P \wedge Q)) \Leftrightarrow ((\neg R \vee P) \wedge (\neg R \vee Q))$ と変形できる。よって、$(R \wedge \neg P)$ と $(R \wedge \neg Q)$ の両方が偽となることを示せば、背理法により $R \Rightarrow P \wedge Q$ が成り立つことを示すことができる。

命題は、$W \wedge P \wedge Q \Leftrightarrow R$ であるが、本稿では書面の都合で $W$ で選ばれるラベルがクラスタ $C_i$ 内の過半数の辺が共通してもつラベルである場合の証明のみ記載する。この場合、$W$ で選ばれるラベルは必ず $P$ でも選ばれるため、$P \wedge W$ と $P$ は同じ意味となる。また、$Q$ は [辺の半数未満しか共通していない、どのラベル $\ell_k$ もクラスタラベル $c\ell(C_i)$ に含まれない] と簡略化できる。

よって、$P \wedge Q \Leftrightarrow R$ の証明を考える。まず、$R \Rightarrow P \wedge Q$ について考える。次に、$P \wedge Q \Rightarrow R$ について考える。

### 4.1 $R \Rightarrow P \wedge Q$ の証明

#### 4.1.1 $(R \wedge \neg P \Leftrightarrow 0)$ の証明

あるクラスタ $C_i$ において、$C_i$ 内の辺の過半数が共通して、もつあるラベル $\ell_j$ がクラスタラベル $c\ell(C_i)$ に含まれないときに、クラスタ $C_i$ の $cost$ は最小であると仮定する。$C_i$ 内の辺の過半数が共通して、もつラベル $\ell_j$ がクラスタラベルに含まれない場合 ($c\ell(C(i))$) の $cost$ と $\ell_j$ を追加した場合 ($c\ell(C(i)) \cup \{\ell_j\}$) の $cost$ との差分 $\Delta\ell_j$ を計算する。差分 $\Delta\ell_j$ は、

$$\Delta\ell_j = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), c\ell(C(i))) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), c\ell(C(i)) \cup \{\ell_j\}) \quad (5)$$

となる。ここで、$C_i$ 内の辺一本に着目すると $(i)\ell(x,y)$ に $\ell_j$ が含まれている場合と $(ii)\ell(x,y)$ に $\ell_j$ が含まれていない場合の差分が考えられ、$(i)$ の差分は、$\ell_j \in \ell(x,y)$ なら、$d_\ell(\ell(x,y), c\ell(C(x))) - d_\ell(\ell(x,y), c\ell(C(x)) \cup \{\ell_j\}) = 1$ で $(ii)$ の差分は $\ell_j \notin \ell(x,y)$ なら、$d_\ell(\ell(x,y), c\ell(C(x))) - d_\ell(\ell(x,y), c\ell(C(x)) \cup \{\ell_j\}) = -1$ となる。そして、$C_i$ 内の $(i)$ と $(ii)$ の差分の合計は、

$$\sum_{(x,y)\in V_2, x,y\in C_i} \Delta\ell_j(x,y) = \sum_{(x,y)\in V_2, x,y\in C_i, \ell_j\in\ell(x,y)} d_\ell(\ell(x,y), c\ell(C(i))) + \sum_{(x,y)\in V_2, x,y\in C_i, \ell_j\notin\ell(x,y)} d_\ell(\ell(x,y), c\ell(C(i)) \cup \{\ell_j\}) \quad (6)$$

となる。ここで、$C_i$ 内の辺の数を $n$、過半数が共通して、もつラベル $\ell_j$ を含んだ辺の数を $n_{\ell_j}$ とすると、式 6 の第一項は、$(i)$ の差分である 1 を $n_{\ell_j}$ 回かけたものなので、$1 \times n_{\ell_j}$ と表せ、$\ell(x,y)$ に $\ell_j$ が含まれていない辺の数を $(n - n_{\ell_j})$ と表現できるので、第二項は、$(ii)$ の差分である $-1$ を $n - n_{\ell_j}$ 回かけたものなので、$-1 \times (n - n_{\ell_j})$ と表せる。つまり、$\Delta\ell_j = \sum_{(x,y)\in V_2, x,y\in C_i} \Delta\ell_j(x,y) = 1 \times n_{\ell_j} + \{-1 \times (n - n_{\ell_j})\}$ である。これを計算すると、$2n_{\ell_j} - n$ となり、$n_{\ell_j}$ が $n$ の過半数であるという仮定より、$n_{\ell_j}$ の 2 倍は、クラスタ $C_i$ 内の辺の数 $n$ より大きくなるので、$2n_{\ell_j} - n > 0$ となる。したがって、あるクラスタ $C_i$ において、$C_i$ 内の辺の過半数が共通して、もつあるラベル $\ell_j$ がクラスタラベル $c\ell(C_i)$ に含まれないときに、$c\ell(C_i)$ に $\ell_j$ を追加することで $cost$ を削減できるため、クラスタ $C_i$ の $cost$ は最小ではない。よって、矛盾が生じるので $R \wedge \neg P \Leftrightarrow 0$ となる。

#### 4.1.2 $(R \wedge \neg Q \Leftrightarrow 0)$ の証明

あるクラスタ $C_i$ において、辺の半数未満しか共通していない、あるラベル $\ell_k$ がクラスタラベル $c\ell(C_i)$ に含まれるときに、クラスタ $C_i$ の $cost$ は最小であると仮定する。$C_i$ 内の辺の半数未満しか共通していないラベル $\ell_k$ がクラスタラベルに含まれる場合の $cost$ と $\ell_k$ をクラスタラベルから取り除いた場合の $cost$ との差分 $\Delta\ell_k$ を計算する。今、少なくとも 1 本は過半数の辺が共通してもつラベルが存在するという仮定より、$c\ell(C_i) - \{\ell_k\} \neq \phi$ が成り立つ。このとき、差分 $\Delta\ell_k$ は、

$$\Delta\ell_k = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), c\ell(C(i))) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), c\ell(C(i)) - \{\ell_k\}) \quad (7)$$

となる。ここで、$C_i$ 内の辺一本に着目すると $(i)\ell(x,y)$ に $\ell_k$ が含まれている場合の差分と $(ii)\ell(x,y)$ に $\ell_k$ が含まれていない場合の差分が考えられ、$(i)$ の差分は $\ell_k \in \ell(x,y)$ なら、$d_\ell(\ell(x,y), c\ell(C(x))) - d_\ell(\ell(x,y), c\ell(C(x)) - \{\ell_k\}) = -1$ で $(ii)$ の差分は $\ell_k \notin \ell(x,y)$ なら、$d_\ell(\ell(x,y), c\ell(C(x))) - d_\ell(\ell(x,y), c\ell(C(x)) - \{\ell_k\}) = 1$ となる。そして、$C_i$ 内の $(i)$ と $(ii)$ の差分の合計は、

$$\sum_{(x,y)\in V_2, x,y\in C_i} \Delta\ell_k(x,y) = \sum_{(x,y)\in V_2, x,y\in C_i, \ell_k\in\ell(x,y)} d_\ell(\ell(x,y), c\ell(C(i))) + \sum_{(x,y)\in V_2, x,y\in C_i, \ell_k\notin\ell(x,y)} d_\ell(\ell(x,y), c\ell(C(i)) - \{\ell_k\}) \quad (8)$$

となる。ここで、$C_i$ 内に含まれる辺の数を $n$、過半数未満しか共通していない、あるラベル $\ell_k$ を含んだ辺の数を $n_{\ell_k}$ とすると、式 8 の第一項は、$(i)$ の差分である $-1$ を $n_{\ell_k}$ 回かけたものなので、$-1\times n_{\ell_k}$ と表せ、$\ell(x,y)$ に $\ell_k$ が含まれていない辺の数を $(n - n_{\ell_k})$ と表現できるので、第二項は、$(ii)$ の差分である 1 を $(n - n_{\ell_k})$ 回かけたものなので、$1\times\ (n - n_{\ell_k})$ と表せる。つまり、$\Delta\ell_k = \sum_{(x,y\in V_2, x,y\in C_i)} \Delta\ell_k(x,y) = -1\times n_{\ell_k} + 1\times(n - n_{\ell_k})$ である。これを計算すると、$n - 2n\ell_k$ となり、$n - 2n\ell_k > 0$ となる。したがって、あるクラスタ $C_i$ において、辺の半数未満しか共通していない、あるラベル $\ell_k$ がクラスタラベル $c\ell(C_i)$ に含まれるときに、$c\ell(C_i)$ から $\ell_k$ を削除することで $cost$ を削減できるため、クラスタ $C_i$ の $cost$ は最小ではない。よって矛盾が生じるので、$R \wedge \neg Q \Leftrightarrow 0$ となる。$(R \wedge \neg P) \vee (R \wedge \neg Q) \Leftrightarrow 0$ が成り立つことを示せたので、$R \Rightarrow P \wedge Q$ が証明できた。

### 4.2 $(P \wedge Q \Rightarrow R)$ の証明

先の $R \Rightarrow P \wedge Q$ が証明できたことにより、系として $P \wedge Q$ の条件を満たすクラスタラベルの 1 つは $cost$ を最小とすることを容易に導くことができる。この系を用いて、$P \wedge Q$ を満たすラベルの集合が 1 つだった場合には、$P \wedge Q \Rightarrow R$ も自明に成り立つ。一方、$P \wedge Q$ を満たすラベルの集合が複数存在する場合には、先の系より $P \wedge Q$ を満たす任意のラベルの集合の組 $(S, S')$ において、$S$ をクラスタラベルとしたときと、$S'$ をクラスタラベルとしたときの $cost$ が同じ値であることを示すことで命題を証明できる。

$P \wedge Q$ を満たす任意のラベルの集合が 2 つ以上存在する場合を考える。$P \wedge Q$ を満たすラベルの集合の任意の異なる二項組 $(S, S')$ を考える。
$\ell_h \in (S - S')$ を満たす任意のラベル $\ell_h$ を考えるとき、今、辺の過半数が共通してもつラベルが存在するという仮定より、$\ell_h$ が満たす性質はクラスタ $C_i$ 内の辺の半数のみに共通していることである。$\ell_h$ がクラスタラベルに含まれる場合の $cost$ と $\ell_h$ をクラスタラベルから取り除いた場合の差分 $\Delta\ell_h$ を計算すると、

$$\Delta\ell_h = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S - \{\ell_h\}) \quad (9)$$

となる。ここで、$C_i$ 内の辺一本に着目すると (I)$\ell(x,y)$ に $\ell_h$ が含まれている場合の差分と (II)$\ell(x,y)$ に $\ell_h$ が含まれていない場合の差分が考えられ、(I) の差分は $\ell_h \in \ell(x,y)$ なら、$d_\ell(\ell(x,y), S) - d_\ell(\ell(x,y), S - \{\ell_h\}) = -1$ で (II) の差分は $\ell_h \notin \ell(x,y)$ なら、$d_\ell(\ell(x,y), S) - d_\ell(\ell(x,y), S - \{\ell_h\}) = 1$ となる。ここで、(I) の差分が-1 と (II) の差分が 1 で $C_i$ 内の $\ell_h$ が含まれている辺と含まれていない辺は同じ本数あるので、$C_i$ 内の差分 $\Delta\ell_h$ は 0 になる。辺の過半数が共通してもつラベルが存在するという仮定より $S - (S - S') \neq \phi$ であり、式 (9) より、

$$\sum_{\ell_h\in(S-S')} \Delta\ell_h = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S - (S - S')) \quad (10)$$

と表せ、$\Delta\ell_h = 0$ より、$\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S - (S - S')) = 0$ となる。よって、

$$\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S) = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S - (S - S')) \quad (11)$$

次に、$\ell_m \in (S' - S)$ を満たす任意のラベル $\ell_m$ を考え

る。このとき、$\ell_m$ が満たす性質もクラスタ $C_i$ 内の辺の半数のみに共通していることである。$\ell_m$ はクラスタラベル集合 $S-(S-S')$ には含まれないので、クラスタラベルに含まれない場合の $cost$ と $\ell_m$ をクラスタラベルに追加した場合の差分 $\Delta\ell_m$ を計算すると、

$$\Delta\ell_m = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S-(S-S')) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), (S-(S-S'))\cup\{\ell_m\}) \quad (12)$$

となる。ここで、$C_i$ 内の辺一本に着目すると (III)$\ell(x,y)$ に $\ell_m$ が含まれている場合の差分と (IV)$\ell(x,y)$ に $\ell_m$ が含まれていない場合の差分が考えられ、(III) の差分は $\ell_m \in \ell(x,y)$ なら、$d_\ell(\ell(x,y), S-(S-S'))-d_\ell(\ell(x,y), S-(S-S')\cup\{\ell_m\}) = -1$ で (IV) の差分は $\ell_m \notin \ell(x,y)$ なら、$d_\ell(\ell(x,y), S-(S-S'))-d_\ell(\ell(x,y), S-(S-S')\cup\{\ell_m\}) = 1$ となる。ここで、(V) の差分が-1 と (VI) の差分が 1 で $C_i$ 内の $\ell_h$ が含まれている辺と含まれていない辺は同じ本数あるので、$C_i$ 内の差分 $\Delta\ell_m$ は 0 になる。式 (12) より、

$$\sum_{\ell_m\in(S'-S)} \Delta\ell_m = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S-(S-S')) - \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), (S-(S-S'))\cup(S'-S)) \quad (13)$$

と表せ、$\Delta\ell_m = 0$ より、$\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S-(S-S'))-\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), (S-(S-S'))\cup(S'-S)) = 0$ となる。よって、

$$\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S-(S-S')) = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), (S-(S-S'))\cup(S'-S)) \quad (14)$$

となる。今、$S$ を加減すると、$(S-(S-S'))\cup(S'-S) = S'$ のように $S'$ へ変形できるので、

$$\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S-(S-S')) = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S') \quad (15)$$

と表せる。式 (11), (15) より

$$\sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S) = \sum_{(x,y)\in V_2, x,y\in C_i} d_\ell(\ell(x,y), S') \quad (16)$$

が導ける。式 (16) より、$P\wedge Q$ を満たす任意の 2 つのラベルの集合 $S$ と集合 $S'$ は、同じ $cost$ をもつので、$S$ をクラスタラベルとしたときの $cost$ である $cost(S)$ と $S'$ をクラスタラベルとしたときの $cost$ である $cost(S')$ は等しい。つまり、任意の異なる二項組に対して成り立つので、$P\wedge Q$ を満たすラベルの集合全てが同じ $cost$ であることが示せた。一方、$R\Rightarrow P\wedge Q$ より、$P\wedge Q$ を満たすクラスタラベルのうち 1 つは必ず $R$ を満たす。このクラスタラベルを $S$ としたとき、$cost(S)$ は最小なので、$cost(S)$ と等しい $cost(S')$ も最小となる。したがって、$P\wedge Q\Rightarrow R$ が証明できたので、 $P\wedge Q\Leftrightarrow R$ が成り立つ。

## 5. まとめ

本稿では辺ラベル付きグラフのクラスタにおいて、各クラスタを表現するラベルの集合を決定するクラスタラベル決定アルゴリズムを提案し、その最適性を証明した。辺色付きグラフのクラスタリングにおいて、クラスタが与えられたときにそのラベルを決定する問題は文献 [1] で辺のラベルが 1 種類の場合である CHROMATIC-CORRELATION-CLUSTERING 問題を対象として考察されていた。文献 [1] では、決定法としてクラスタ内のクラスタラベルで多数決を取り、1 つを選択するというものであったが、最適性の詳細な証明はなかった。本研究では、上記の決定法を多種類の辺ラベルをもつグラフのクラスタのラベルを決定する方法に拡張し、文献 [1] の決定法を含め、その最適性を詳細に証明した。

今後の課題としては、提案クラスタラベル決定法を基とした、トップダウンなクラスタリング方法の提案が挙げられる。

### 参考文献

[1] Francesco Bonchi, Aristides Gionis, Francesco Gullo, Charalampos E. Tsourakakis, and Antti Ukkonen, “ Chromatic correlation clustering, ” ACM Trans. Knowledge Discovery from Data, Vol.9, No.4, Article 34, 2015.
[2] Wenfei Fan, Jianzhong Li, Shuai Ma, Nam Tang, and Yinghui Wu,”Adding regular expressions to graph reachability and pattern gueries,” Proc. of IEEE Internationl Conf. on Data Engineering (ICDE’11), pp.39-50, 2011.
[3] Ruoming Jin, Hui Hong, Haixum Wang, Ning Ruan, and Yang Xiang, ”Computing label-constraint reachability in graph databases,” Proc. of ACM SIGMOD International Conf. on Management of Data(SIGMOD’10), pp.123-134, 2010.
[4] Kun Xu, Lei Zou, Jeffery Xu Yu, Lei Chen, Yanghua Xiao, and Dongyan Zhao, ”Answering label-constraint reachability in large graphs,” Proc. of ACM International Conf. on Information and Knowledge Management(CIKM’11), pp.1595-1600, 2011.