数理解析研究所講究録
第1773巻 2012年 121-131

隣接数に着目したハイパーグラフ上のコミュニティ抽出

筑波大学大学院　システム情報工学研究科,
宮川裕幸 (Hiroyuki Miyagawa), 繁野麻衣子 (Maiko Shigeno),
高橋里司 (Satoshi Takahashi), 張明超 (Mingchao Zhang)
Graduate School of Systems and Information Engineering,
University of Tsukuba

概要　ウェブグラフや社会ネットワークにおけるコミュニティ抽出はネットワーク分析で重要な役割を果たしている．コミュニティは，対象となるネットワークをグラフとして表現したとき，密な部分グラフとして定義され，様々な抽出アルゴリズムが研究されてきた．一方で，複雑なネットワークを表現するのに，いくつかのノードの集まりや属性の違いなどを表現するハイパーグラフが注目されている．本研究ではグラフ上で隣接数を基に定義されたコミュニティをハイパーグラフ上に拡張し，4つの拡張モデルを定義する．そして，いずれのモデルに対しても最小カットアルゴリズムを利用した抽出アルゴリズムが構築できることを示す．また，共著データを用いてモデルの違いによる抽出コミュニティの相違を検証する．

キーワード：コミュニティ抽出; ハイパーグラフ; 最小カットアルゴリズム

1　はじめに

ウェブグラフや社会ネットワークにおけるコミュニティ抽出はネットワーク分析で重要な役割を果たしている．対象となるネットワークの中で共通の関心や利害関係をもつ集団をコミュニティという．ウェブグラフでは，関連があるトピックを持つウェブページの集まりをコミュニティと呼ぶ．ウェブリンクは関連あるページ間に張られるので，ウェブリンクが密なウェブページの集まりはコミュニティと見なすことができる．このように，対象となるネットワークをグラフで表現すると，コミュニティをなすノードの集合は，そのノードが誘導する部分グラフ内ではエッジが密であり，かつ，コミュニティをなすノードとそれ以外のノードの間のエッジが疎となる．コミュニティ抽出に関する研究では，このエッジの疎密に対して様々な角度から定義を与え，抽出アルゴリズムが構築されてきた．

コミュニティの抽出手法は大きく2種類に分けられる．一つは一度に1個あるいは少数のコミュニティを抽出する方法であり，もう一つはすべてのノードをあるコミュニティに属するように，ネットワーク全体をコミュニティに分割する方法である．コミュニティ抽出では後者が一般的な方法である．しかし，ネットワークをコミュニティに分割すると，関係の強くないノードが同じコミュニティに分割される可能性があり，適用する対象によっては好ましくないこともある．本研究は一度に1個あるいは少数のコミュニティを抽出する方法に注目する．グラフ構造に着目した古典的な方法に，クリークの条件を緩和した部分グラフを抽出するものがある．Flake-Lawrence-Giles[4] はノード間の隣接数によりコミュニティを定義し，最小カットアルゴリズムを用いて効率的にコミュニティを抽出している．彼らの提案したコミュニティは隣接数に着目していることから，本論文では adj-コミュニティ (adjacent number community) と呼ぶ．adj-コミュニティは局所的な情報のみから判断できるので，大規模な

ネットワーク上でも効率よく抽出できるコミュニティとして知られている.

一方, ノード間の関係が複数ノードの小グループによって与えられるようなネットワークを表現するのには, グラフよりもハイパーグラフが適している. 近年, ハイパーグラフ上のコミュニティ抽出の研究も盛んである. Brinkmeier-Recknagel-Werner [3] は, ハイパーグラフ上の最小カットによりコミュニティを定義し, 指定した集合を含むコミュニティの抽出を行っている. 張-高橋-繁野 [9] はハイパーグラフ上のハイパーエッジ数とノード数の比率によりコミュニティを定義し, 効率のよい抽出アルゴリズムを提案している. Barber [2] は, ネットワークのノードと属性を 2 部グラフで表現し, モジュラリティによりコミュニティを抽出する方法を提案しており, さまざまなアルゴリズムが派生している.

本研究では, グラフ上で定義されている adj-コミュニティをハイパーグラフ上に拡張し, 4 つのコミュニティモデルを定義して, これらのコミュニティを抽出するアルゴリズムを構築する. そして, 実際の共著データを用いて 4 つの拡張モデルによる抽出コミュニティの相違を比較する. いずれのモデルも最小カットアルゴリズムにより効率よくコミュニティを抽出でき, 実問題への適用時に, モデルの拡張性の指針となることが期待される. 本研究では, 抽出コミュニティの相違の検証を目的とするために, ハイパーグラフはすべてデータとして与えられており, 静的なハイパーグラフを扱う.

2 準備

ノード集合を N, エッジ集合を E とする無向グラフ $G=(N,E)$ において, ノード $v \in N$ と ノードの部分集合 $C \subseteq N$ に対して, v と C の間のエッジ集合を

$$\delta_G(v,C) = \{(v,w) \in E \mid w \in C\}$$

とする. Flake-Lawrence-Giles[4] は, コミュニティの内部の関係が外部よりも密であることに着目したコミュニティの定義を以下のように与えている.

定義 1 ([4]) 無向グラフ $G=(N,E)$ において, ノードの真部分集合 $C(\subset N)$ の各ノード $v(\in C)$ で, $|\delta(v,C)| \geq |\delta(v,N \setminus C)|$ を満たすとき, C を **adj-コミュニティ**という.

グラフ G のノード s と t に対し, ノード集合の分割 $(X, N \setminus X)$ が $s \in X, t \in N \setminus X$ を満たすとき, $(X, N \setminus X)$ を s-t カットという. s-t カット $(X, N \setminus X)$ の中で, X と $N \setminus X$ を結ぶエッジ数 $|\{(u,v) \in E \mid u \in X, v \in N \setminus X\}|$ を最小とする s-t カットを最小 s-t カットという. グラフ G のエッジが容量 $c: E \to \mathbb{R}$ をもつときは, s-t カット $(X, N \setminus X)$ の中で, X と $N \setminus X$ の間のエッジの容量の和であるカット容量 $c(X, N \setminus X)$, すなわち,

$$c(X, N \setminus X) = \sum \{c(u,v) \mid u \in X, v \in N \setminus X\}$$

を最小とする s-t カットを最小 s-t カットという. 有向グラフにおいては, $c(X, N \setminus X)$ は X から $N \setminus X$ への向きのエッジの容量の和で与える. 最小 s-t カットを見つける効率的な多項式時間アルゴリズムは数多く研究されている [1]. 以下の性質は, adj-コミュニティは無向グラフの最小 s-t カットアルゴリズムを用いて効率的に見つけることができることを示している.

定理 1　ノード s と t に対する最小 s-t カットを $(X, N \setminus X)$ とすると，

$$|\delta_G(s, X)| \geq |\delta_G(s, N \setminus X)| \tag{1}$$

を満たせば，X は adj-コミュニティである．　□

そこで，グラフ G が与えられれば，adj-コミュニティは以下の手続きを繰り返すことで見つけられる．

(ステップ 1)　ノード s と t を選ぶ．
(ステップ 2)　最小 s-t カット $(X, N \setminus X)$ を求める．
(ステップ 3)　求めた X が条件 (1) を満たすとき，X は adj-コミュニティであり，X を出力．

Ino-Kudo-Nakamura [6] は，この抽出では s を含み t を含まないすべてのコミュニティを抽出することができないのみならず，コミュニティが存在していても，抽出できない可能があることを指摘している．しかし，Flake-Lawrence-Giles [4] はウェブグラフ上の実問題に対して，ウェブグラフの特性を用いてノード s と t を選ぶことで，その実用性を示している．特に，グラフ全体が分からなくても，決まった深さまでエッジを探索することで得られるグラフ上に適用できる利点がある．また，adj-コミュニティから派生したコミュニティもある [6]．すなわち，adj-コミュニティはコミュニティ抽出に対する基本的な概念の一つである．そこで，本研究では adj-コミュニティに着目し，ハイパーグラフへの拡張を試みる．

3　ハイパーグラフ上のコミュニティ

与えられた有限集合 N に対し，$\mathcal{P}^*(N) = \{X \subseteq N \mid |X| \geq 2\}$ を要素が少なくとも 2 である N のすべての部分集合とする．有限なノード集合 N と $\mathcal{H} \subseteq \mathcal{P}^*(N)$ からなるハイパーグラフを $\Gamma = (N, \mathcal{H})$ と書く．$\mathcal{H}$ の要素をハイパーエッジという．すべてのハイパーエッジの要素数が 2 であるときのハイパーグラフは（無向）グラフとなる．本節ではグラフ G 上で定義された adj-コミュニティをハイパーグラフ $\Gamma = (N, \mathcal{H})$ 上に拡張する．

グラフ G のノード $v \in N$ とノードの部分集合 $C \subseteq N$ の間のエッジ集合 $\delta_G(v, C)$ の要素数 $|\delta_G(v, C)|$ を隣接エッジ数という．これに対応して，ハイパーグラフ Γ では，ノード $v \in N$ に対し，v を含むハイパーエッジの集合を

$$\delta(v) = \{h \in \mathcal{H} \mid v \in h\}$$

とし，ノード $v \in N$ とノードの部分集合 $C \subseteq N$ を関連づけるハイパーエッジの集合を

$$\delta_\Gamma(v, C) = \{h \in \delta(v) \mid h \subseteq C \cup \{v\}\}$$

とする．この要素数 $|\delta_\Gamma(v, C)|$ を隣接ハイパーエッジ数とよぶ．ここで，Γ のすべてのハイパーエッジの要素数が 2 であるとき $\delta_G(v, C) = \delta_\Gamma(v, C)$ であるので，隣接ハイパーエッジ数は隣接エッジ数の自然な拡張である．隣接ハイパーエッジ数を用いて，adj-コミュニティをハイパーグラフ上に拡張する．

定義 2　ハイパーグラフ $\Gamma=(N,\mathcal{H})$ において，ノードの真部分集合 $C(\subset N)$ の各ノード $v\in C$ で，

$$|\delta_\Gamma(v,C)|\geq|\delta_\Gamma(v,N\setminus C)| \tag{2}$$

を満たすとき，C を **h-コミュニティ** (hyperedge-based-community) という．

h-コミュニティ抽出のために容量つきの有向グラフ $(\tilde{D}_\Gamma,\tilde{c})$ を作成する．ハイパーエッジ $h\in\mathcal{H}$ の 2 つのコピーをそれぞれ h^+,h^- と表し，$\mathcal{H}^+=\{h^+\mid h\in\mathcal{H}\}$, $\mathcal{H}^-=\{h^-\mid h\in\mathcal{H}\}$ とする．$\tilde{D}_\Gamma$ は $\tilde{N}=N\cup\mathcal{H}^+\cup\mathcal{H}^-$ をノード集合とし，$\tilde{A}=\{(v,h^+),(h^-,v)\mid v\in N,h\in\delta(v)\}$ と $\tilde{A}^\pm=\{(h^+,h^-)(h^-,h^+)\mid h\in\mathcal{H}\}$ の和集合をエッジ集合とする有向グラフであり，エッジの容量 $c:\tilde{A}\cup\tilde{A}^\pm\to\mathbb{R}$ は

$$\tilde{c}(a)=\begin{cases}\infty & (a\in\tilde{A}),\\ 1 & (a\in\tilde{A}^\pm).\end{cases}$$

で与える．

定理 2　容量付き有向グラフ $(\tilde{D}_\Gamma,\tilde{c})$ 上で，ノード $s,t\in N$ に対する最小 s-t カット $(Y,\tilde{N}\setminus Y)$ において，$C=Y\cap N$ とする．$|\delta_\Gamma(s,C)|\geq|\delta_\Gamma(s,N\setminus C)|$ ならば，C は h-コミュニティである．

証明. 最小 s-t カットは必ず有限なカット容量をもつ．なぜならば，例えば，$\tilde{c}(N\cup\mathcal{H}^+,\mathcal{H}^-)=|\mathcal{H}|$ のようにカット容量が有限な s-t カットが存在するからである．

$\hat{v}\in C$ と仮定する．もし，$\hat{v}$ を含むハイパーエッジ $h(\in\delta(\hat{v}))$ が $h^+\notin Y$ ならば，カット容量 $\tilde{c}(Y,\tilde{N}\setminus Y)\geq\tilde{c}(v,h^+)=\infty$ となりこれは最小 s-t カットではない．よって $\{h^+\mid h\in\delta(\hat{v})\}\subseteq Y$ であり，それ故 $\{h^+\mid h\in\delta_\Gamma(\hat{v},C)\}$ と $\{h^+\mid\delta_\Gamma(\hat{v},N\setminus C)\}$ はどちらも Y に含まれる．同様に，$h^-\in Y$ ならば h に含まれるいずれのノード v も C になければならない．よって，$\{h^-\mid h\in\delta_\Gamma(\hat{v},N\setminus C)\}\subseteq\tilde{N}\setminus Y$ を満たす．さらに，$h\in\delta_\Gamma(\hat{v},C)$ に対し，$h^-\notin Y$ ならば，$\tilde{c}(Y,\tilde{N}\setminus Y)=\tilde{c}(Y\setminus\{h^-\},\tilde{N}\setminus(Y\setminus\{h^-\}))+1$ となるので，$\tilde{c}(Y,\tilde{N}\setminus Y)$ が最小 s-t カットであることと矛盾する．以上より，$\{h^-\mid h\in\delta_\Gamma(\hat{v},C)\}\subseteq Y$ を得る．

ここで，$|\delta_\Gamma(v,C)|<|\delta_\Gamma(v,N\setminus C)|$ である $v\in C\setminus\{s\}$ が存在すると仮定する．すると，$Y'=Y\setminus(\{v\}\cup\{h^-\mid h\in\delta_\Gamma(v,C)\}\cup\{h^+\mid h\in\delta_\Gamma(v,N\setminus C)\})$ とおくと，$(Y',\tilde{N}\setminus Y')$ は s-t カットであり，

$$\tilde{c}(Y',\tilde{N}\setminus Y')=\tilde{c}(Y,\tilde{N}\setminus Y)+|\delta_\Gamma(v,C)|-|\delta_\Gamma(v,N\setminus C)|<\tilde{c}(Y,\tilde{N}\setminus Y)$$

となり，これは $(Y,\tilde{N}\setminus Y)$ が最小 s-t カットであることと矛盾する．従って，任意の $v\in C\setminus\{s\}$ で条件 (2) が満たされており，s に対して条件 (2) が満たされさえすれば h-コミュニティになる．　□

よって，容量付き有向グラフ $(\tilde{D}_\Gamma,\tilde{c})$ を作成すれば，adj-コミュニティの抽出と同様の方法で，ノード $s,t\in N$ を選択後，最小 s-t カットを求めることで h-コミュニティを抽出できる．

隣接エッジ数を隣接ハイパーエッジ数に置き換えた h-コミュニティは，adj-コミュニティの自然な拡張とみなせる．しかし，adj-コミュニティではノード v と接続するすべてのエッジが考慮されていたが，h-コミュニティではノード v を含むすべてのハイパーエッジ $\delta(v)$ が考慮されていない．つまり，$h\in\delta(v)$ で，$h\cap(C\setminus\{v\})\neq\emptyset$ かつ $h\setminus C\neq\emptyset$ である h はコミュニティ判定の条件で考慮されない．

そこで次に，すべてのハイパーエッジを，ノードの部分集合 $C(\subset N)$ の関連性への貢献度で分類する．ノードの真部分集合 $C(\subset N)$ に対し，

$$\mathcal{H}(C)^{>} = \{h \in \mathcal{H} \mid |h \cap C| > |h \setminus C|\}$$
$$\mathcal{H}(C)^{\leq} = \{h \in \mathcal{H} \mid |h \cap C| \leq |h \setminus C|\}$$

とする．

定義 3　ハイパーグラフ $\Gamma = (N, \mathcal{H})$ において，ノードの真部分集合 $C(\subset N)$ の各ノード $v \in C$ で，

$$|\delta(v) \cap \mathcal{H}(C)^{>}| \geq |\delta(v) \cap \mathcal{H}(C)^{\leq}| \tag{3}$$

を満たすとき，C を **c-コミュニティ** (classified-hyperedges community) という．

すべてのハイパーエッジの要素数が 2 であるとき，つまり，Γ をグラフ G と見なせるとき，$v \in C$ に対して明らかに，$\delta_G(v, C) = \delta(v) \cap \mathcal{H}(C)^{>}$，$\delta_G(v, N \setminus C) = \delta(v) \cap \mathcal{H}(C)^{\leq}$ が成り立つ．よって，c-コミュニティも adj-コミュニティを特殊ケースとして含む．

c-コミュニティを抽出するため 2 部グラフ $\tilde{G}_\Gamma = (N \cup \mathcal{H}, \tilde{E})$ を作成する．ここで $\tilde{E} = \{(v, h) \mid h \in \delta(v), v \in N\}$ とする．

定理 3　2 部グラフ $\tilde{G}_\Gamma$ 上で，ノード $s, t \in N$ に対する最小 s-t カット $(Y, (N \cup \mathcal{H}) \setminus Y)$ において，$C = Y \cap N$ とする，$|\delta(s) \cap \mathcal{H}(C)^{>}| \geq |\delta(s) \cap \mathcal{H}(C)^{\leq}|$ ならば，C は c-コミュニティである．

証明. s-t カット $(X, (N \cup \mathcal{H}) \setminus X)$ に対して，X と $(N \cup \mathcal{H}) \setminus X$ 間のエッジ数を $\Delta(X, (N \cup \mathcal{H}) \setminus X)$ で表す．また，$\mathcal{H}(C)^{=} = \{h \in \mathcal{H} \mid |h \cap C| = |h \setminus C|\}$ とする．s-t カットの中で，$(Y, (N \cup \mathcal{H}) \setminus Y)$ は $\Delta(Y, (N \cup \mathcal{H}) \setminus Y)$ が最小であるので，$\mathcal{H} \cap Y$ のハイパーエッジは $\mathcal{H}(C)^{>}$ あるいは $\mathcal{H}(C)^{=}$ に含まれる．同様に，$\mathcal{H} \setminus Y$ のハイパーエッジは $\mathcal{H}(C)^{\leq}$ に含まれる．

ここで，$|\delta(v) \cap \mathcal{H}(C)^{>}| < |\delta(v) \cap \mathcal{H}(C)^{\leq}|$ を満たす $v \in C \setminus \{s\}$ が存在すると仮定する．すると，$Y' = Y \setminus (\{v\} \cup (\delta(v) \cap \mathcal{H}(C)^{=}))$ とおくと，$(Y', (N \cup \mathcal{H}) \setminus Y')$ は s-t カットであり，

$$\begin{aligned}
&\Delta(Y', (N \cup \mathcal{H}) \setminus Y') \\
&= \Delta(Y, (N \cup \mathcal{H}) \setminus Y) + |\delta(v) \cap \mathcal{H}(C)^{>}| - |\delta(v) \setminus Y| - |\delta(v) \cap \mathcal{H}(C)^{=} \cap Y| \\
&= \Delta(Y, (N \cup \mathcal{H}) \setminus Y) + |\delta(v) \cap \mathcal{H}(C)^{>}| - |\delta(v) \cap \mathcal{H}(C)^{\leq}| < \Delta(Y, (N \cup \mathcal{H}) \setminus Y)
\end{aligned}$$

となり，これは $\Delta(Y, (N \cup \mathcal{H}) \setminus Y)$ が最小であることと矛盾する．従って，任意の $v \in C \setminus \{s\}$ で条件 (3) が満たされており，s に対して条件 (3) が満たされさえすれば c-コミュニティになる．□

よって，グラフ $\tilde{G}_\Gamma$ に対する最小 s-t カットアルゴリズムを利用し，c-コミュニティを抽出することができる．

ここで，もう一度 adj-コミュニティの定義を見直す．adj-コミュニティでは隣接エッジ数 $|\delta_G(v, C)|$ に着目していたが，グラフにおいては，v と C 間の隣接エッジ数は，v と隣接する C に属するノード数に等しい．そこで，次に隣接ノード数から adj-コミュニティを拡張する．ハイパーグラフ上で，ノード v, w を含むハイパーエッジ $h(\in \mathcal{H})$ が存在するとき，v と w は隣接しているとよぶ．ノードの真部分集合 $C(\subset N)$ の中でノード v とハイパーエッジ $h \in \delta(v)$ によって隣接するノードの集合

$\{w \in C \setminus \{v\} \mid w \in h\}$ を $\hat{d}_h(v,C)$ と表し，多重集合 $\bigcup_{h\in\delta(v)} \hat{d}_h(v,C)$ を $\hat{d}(v,C)$ とする．すなわち，v に隣接する C 中のノード w に対して，v と w の両方を含むハイパーエッジが k 本あるとき，$\hat{d}(v,C)$ には w が k 個含まれているとする．

定義 4　ハイパーグラフ $\Gamma = (N, \mathcal{H})$ において，ノードの真部分集合 $C(\subset N)$ の各ノード $v \in C$ で

$$|\hat{d}(v,C)| \geq |\hat{d}(v, N \setminus C)|$$

を満たすとき，C を **n-コミュニティ** (node-based-community) という．

隣接するノード数は $|\hat{d}(v,C)| = \sum_{h\in\delta(v)} |h \cap (C \setminus \{v\})|$ なので，n-コミュニティは各ハイパーエッジを完全グラフで表した，多重エッジをもつ無向グラフ上の adj-コミュニティである．よって，前節で示した adj-コミュニティを抽出する手続きで求めることができる．

次節の数値実験でも示すように，サイズの大きいハイパーエッジに含まれるノードは，そのハイパーエッジを表すエッジが密になるので，n-コミュニティのノードになりやすい傾向がある．この点を改善するため，隣接するノード数と隣接するハイパーエッジ数を組み合わせてコミュニティを定義する．集合 $C(\subset N)$ とノード $v \in C$ の間の関係の強さは多重集合 $d(v,C) = \bigcup_{h\in\delta_\Gamma(v,C)} \hat{d}_h(v,C)$ の要素数で評価し，v と $N \setminus C$ の間の関係の強さは $\delta(v) \setminus \delta_\Gamma(v,C)$ の要素数で評価する．すなわち，コミュニティの内外で異なる指標を用いる．

定義 5　ハイパーグラフ $\Gamma = (N, \mathcal{H})$ において，ノードの真部分集合 $C(\subset N)$ の各ノード $v \in C$ で

$$|d(v,C)| \geq |\{h \in \delta(v) \mid h \not\subseteq C\}| \tag{4}$$

を満たすとき，C を **mc-コミュニティ** (mixed-criterion-community) という．

mc-コミュニティも特殊ケースとして adj-コミュニティを含む．mc-コミュニティを抽出するために容量付きの有向 2 部グラフ $D_\Gamma = (N \cup \mathcal{H}, A^F \cup A^B)$ を作る．ただし，$A^F = \{(v,h) \mid v \in N, h \in \delta(v)\}$，$A^B = \{(h,v) \mid v \in N, h \in \delta(v)\}$ である．エッジの容量 $c : A^F \cup A^B \to \mathbb{R}$ は

$$c(e) = \begin{cases} 1 & (e \in A^F), \\ \infty & (e \in A^B). \end{cases}$$

で与える．

定理 4　容量付きの有向 2 部グラフ (D_Γ, c) 上で，ノード $s, t \in N$ に対するの最小 s-t カット $(Y, (N \cup \mathcal{H}) \setminus Y)$ において，$C = Y \cap N$ とする．$|d(s,C)| \geq |\{h \in \delta(s) \mid h \not\subseteq C\}|$ ならば，C は mc-コミュニティである．

証明．容量の定義により，ハイパーエッジ h が C に含まれる必要十分条件は $h \in Y$ である．ここで，$|d(v,C)| < |\{h \in \delta(v) \mid h \not\subseteq C\}|$ を満たす $v \in C \setminus \{s\}$ が存在すると仮定する．すると，$Y' = Y \setminus (\{v\} \cup \delta_\Gamma(v,C))$ とおくと，$(Y', (N \cup \mathcal{H}) \setminus Y')$ は s-t カットであり，

$$\begin{aligned} c(Y', (N \cup \mathcal{H}) \setminus Y') &= c(Y, (N \cup \mathcal{H}) \setminus Y) + \sum_{h\in\delta(v,C)} |h \setminus \{v\}| - |\{h \in \delta(v) \mid h \not\subseteq C\}| \\ &< c(Y, (N \cup \mathcal{H}) \setminus Y) \end{aligned}$$

となり，$c(Y,(N\cup\mathcal{H})\setminus Y)$ の最小性に反する．従って，任意の $v\in C\setminus\{s\}$ で条件 (4) が満たされており，s に対して条件 (4) が満たされさえすれば mc-コミュニティになる． □

4 数値実験

前節で定義した 4 つのコミュニティの妥当性を検証するために，共著関係を表現したハイパーグラフからコミュニティを抽出する．ハイパーグラフのノードは対象とする論文誌などに掲載されている論文の著者とする．ただし，単著の論文のみ掲載されている著者は除外する．ハイパーエッジは共著の論文に対応し，その論文の著者集合からなる．同じ著者による複数の論文は異なるハイパーエッジとして扱う．すなわち，ハイパーエッジ集合は多重集合で与える．

コミュニティ抽出では，指定した著者を含むなるべく小さなコミュニティを抽出する．以下は h-コミュニティを抽出する手続きである．

(ステップ 1) 抽出したいコミュニティの中心となる著者 s を選ぶ．

(ステップ 2) 容量付き有向グラフ $(\tilde{D}_\Gamma,\tilde{c})$ 上で，各ノード $v\in N\setminus\{s\}$ に対して，最小 s-v カット $(Y_v,\tilde{N}\setminus Y_v)$ を求める．$\min\{\tilde{c}(Y_v,\tilde{N}\setminus Y_v)\mid v\in N\setminus\{s\},|\delta_\Gamma(s,Y_v\cap N)|\geq|\delta_\Gamma(s,N\setminus Y_v)|\}$ を達成する s-v カットを $(Y,\tilde{N}\setminus Y)$ とする．

(ステップ 3) $\min\{\tilde{c}(Y_v,\tilde{N}\setminus Y_v)\mid v\in(Y\cap N)\setminus\{s\}\}$ を達成するノード v を新たに選ぶ．

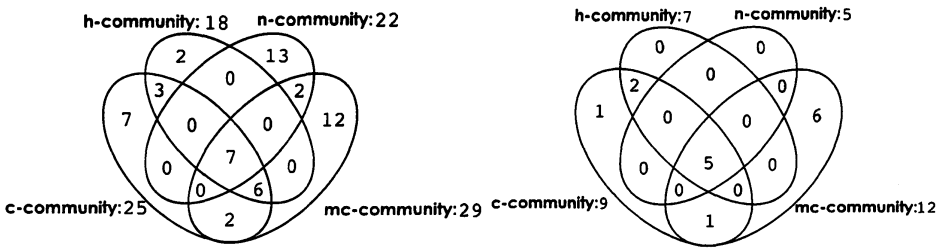
(ステップ 4) $C=(Y\cap Y_v)\cap N$ が条件 $|\delta_\Gamma(s,C)|\geq|\delta_\Gamma(s,N\setminus C)|$ を満たすならば，$(Y\cap Y_v,\tilde{N}\setminus(Y\cap Y_v))$ を新たに $(Y,\tilde{N}\setminus Y)$ と更新してステップ 3 へ．そうでないときは，$Y\cap N$ を h-コミュニティとして出力．

もし，ステップ 2 で $|\delta_\Gamma(s,Y_v\cap N)|\geq|\delta_\Gamma(s,N\setminus Y_v)|$ を満たす s-v カット $(Y_v,\tilde{N}\setminus Y_v)$ が存在しないときは，アルゴリズムはコミュニティを抽出せずに終了する．

同様の手続きで，最小 s-v カットを求めるグラフを変換することで，c-コミュニティ，n-コミュニティ，mc-コミュニティを求める．

数値実験は 2 種類の論文誌の論文共著データと Pajek datasets[7] の “Graph products” による共著データに対して行う．最初に Journal of the Operations Research Society of Japan の 1995 年 38 巻から 2009 年 56 巻の掲載論文のデータを用いる．このとき，ハイパーグラフのノード数は 570 でありハイパーエッジ数は 344 である．コミュニティの中心となる著者として Author1, Author2 を選ぶ．Author1 に接続するハイパーエッジは 9 本であり，Author2 に接続するハイパーエッジは 6 本である．Author1 に対して抽出された h-, c-, n-, mc-コミュニティに含まれる著者数はそれぞれ 18, 25, 21, 29 となった．図 1 (a) のベン図は Author1 に関するそれぞれのコミュニティに属する著者の人数を表す．同様に，図 1 (b) は Author2 に対する結果を示す．Author2 に関して抽出されたコミュニティの関係は，図 2 に 2 部グラフ $\tilde{G}_\Gamma=(N\cup\mathcal{H},\tilde{E})$ で示す．

次に，Mathematical Programming Journal の Series A と Series B の 1989 年 43 巻から 2008 年 115 巻 10 号までの掲載論文のデータを用いる．このとき，ハイパーグラフのノード数は 1800 でありハイパーエッジ数は 1215 である．最初の実験と同様の著者 Author2 に関するコミュニティと別の著



(a) Author1 に関する結果　　(b) Author2 に関する結果

図 1　抽出された各コミュニティに属する著者数（実験 1）

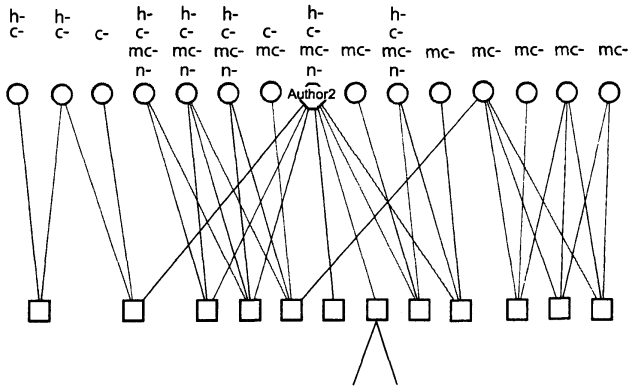
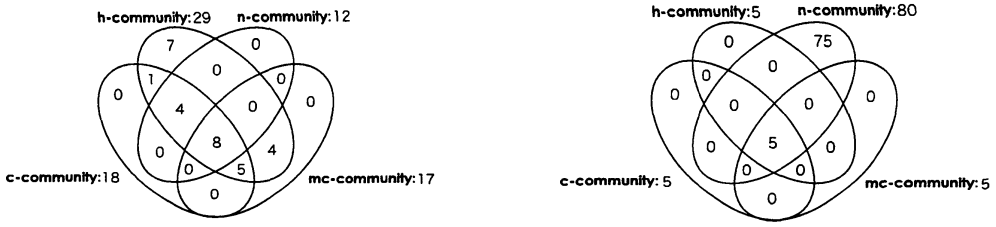


図 2　Author2 に対する 2 部グラフ $\tilde{G}_\Gamma = (N \cup \mathcal{H}, \tilde{E})$ によるコミュニティの関係（実験 1）:丸いノードは著者を表し，四角いノードは論文を表す．ノードの上にそれぞれ属しているコミュニティの種類を表している．コミュニティに属していない著者を含む論文が 1 本ある．

者 Author3 に関するコミュニティを抽出する．Author2, Author3 ともに接続するハイパーエッジは 17 本である．図 3 に Author2, Author3 に関して抽出されたコミュニティに属する著者の人数を示す．Author2 に関しては，抽出されたコミュニティの関係を 2 部グラフ $\tilde{G}_\Gamma = (N \cup \mathcal{H}, \tilde{E})$ で表したものを図 4 に示す．



(a) Author2 に関する結果　　(b) Author3 に関する結果

図 3　抽出された各コミュニティに属する著者数（実験 2）

これら抽出されたコミュニティはいずれも当該分野で活発に研究をしている著者のグループであった．しかし，Author1 の結果にみられるように，n-コミュニティはサイズの大きなハイパーエッジに影響されやすい．実際，6 人の共著論文に対応するハイパーエッジは他に比べてサイズが大きく，n-コ

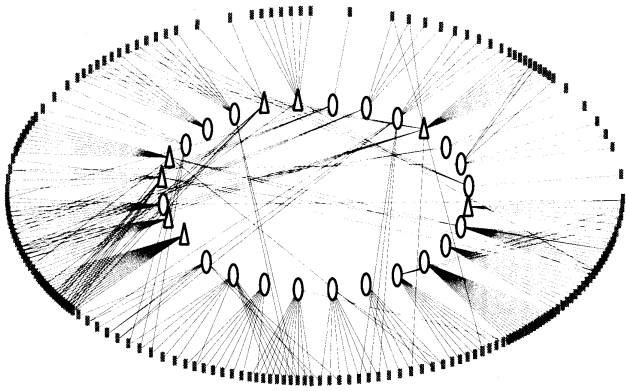


図4　Author2 に対する 2 部グラフ $\tilde{G}_\Gamma = (N \cup \mathcal{H}, \tilde{E})$ によるコミュニティの関係（実験 2）:丸と三角形のノードは著者を表し，外側の四角いノードは論文を表す．抽出されたコミュニティに属する著者はいずれも h-コミュニティに属している．4 つのコミュニティすべてに属する著者 8 人を三角形で示す．

ミュニティにはこの 6 名が含まれていた．Author3 に関しても同様の傾向がみられた．一方，Author2 に関しては，関係するハイパーエッジのサイズが 4 以下であり，得られた n-コミュニティはいずれも他のコミュニティと大きな違いはなかった．また，図 2 からも分かるように，より小さいコミュニティを抽出する手続きにも関わらず，極小なコミュニティが抽出されるわけではなかった．

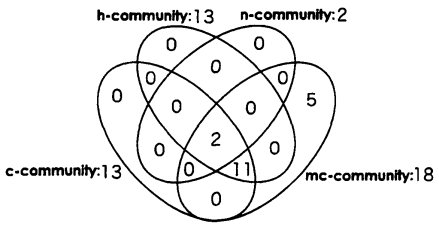
これら抽出コミュニティと文献 [9] による抽出コミュニティを比較する．文献 [9] では，ハイパーグラフ上の最大密度部分集合を求め，それをコミュニティとして定義している．ハイパーグラフ $\Gamma = (N, \mathcal{H})$ に対して，$\Gamma(S) = \{J \in \mathcal{H} \mid J \subseteq S\}$ とする．このとき，

$$\max_{S \subset N, S \neq \emptyset} \frac{|\Gamma(S)|}{|S|}$$

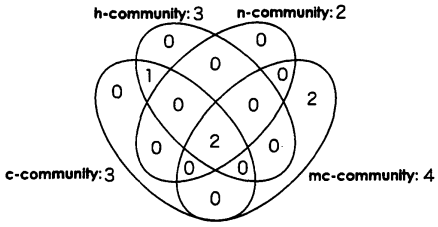
を達成する S を最大密度部分集合といい，最大密度部分集合を求める問題を最大密度部分集合問題という．文献 [9] では，今回の実験と同じ Mathematical Programming Journal の論文共著データを使用している．ただし，最大密度部分集合によるコミュニティでは，特定の中心を指定せず，1 つのコミュニティを抽出する．[9] の実験では，Author2 を含むコミュニティが抽出された．抽出されたコミュニティは，30 人の著者が含まれている．今回の実験結果との比較をすると，h-コミュニティと共通する著者は 18 人，c-コミュニティと共通する著者は 12 人，n-コミュニティと共通する著者は 10 人，mc-コミュニティと共通する著者は 12 人であった．今回の実験では，どのコミュニティも最大密度部分集合によるコミュニティと共通する著者を 1/3 以上抽出している事がわかる．さらに，今回の 4 つのコミュニティに共通する全ての著者は最大密度部分集合によるコミュニティに属していることが確かめられた．

次に，Pajek detests のデータ “Graph products” を用いてコミュニティ抽出を行う．“Graph products” は，W. Imrich と S. Klavžar の著作 [5] での参考文献として引用されている文献の共著データである．データは，ノード数が 314，ハイパーエッジ数が 613 のハイパーグラフである．コミュニティの中心となる著者として，[5] の著者である，W. Imrich と S. Klavžar を選び，それぞれを中心としたコミュニティを抽出する．W. Imrich が著者として入っている文献数は 26 本で，S. Klavžar が著者とし

て入っている文献数は 22 本であり，2 人に共通する文献は 16 本である．図 5 にそれぞれの著者を中心とするコミュニティ抽出結果を示す．図 5 を見ると，h-コミュニティと c-コミュニティは同一の著者



(a)W. Imrich を中心とするコミュニティ



(b)S. Klavžar を中心とするコミュニティ

図 5　抽出された各コミュニティに属する著者数

を抽出していることがわかる．全てのコミュニティに共通する著者 2 人は，どちらも W. Imrich と S. Klavžar である．(a)，(b) どちらも n-コミュニティでは，この 2 人のみを抽出している．これは，この 2 人の著者のみで書かれている参考文献数が 10 本と多いからであると考えられる．さらに，(a) での mc-コミュニティは，他のどのコミュニティにも属さない著者が 5 人いるが，それらは，中心とした著者との共著ではなく，h-や c-コミュニティに属す著者との共著が多い事が特徴である．また，参考論文数が 10 本以上の著者が多い．mc-コミュニティでは，著者の隣接関係とハイパーエッジの本数で特徴付けられているため，各コミュニティに属す著者の周りの著者が一緒に引きずられた可能性がある．(a) と (b) でコミュニティに含まれる著者の数に差が見られるが，これは，各著者の参考文献数の差であると考えられる．

実験を通して，各コミュニティ間の相違を確認できた．ハイパーグラフでモデル化できるネットワークは実社会において少なくはない．今回は共著データを実験対象としてある程度妥当なコミュニティを抽出できたと言える．実用に際しては，対象とするデータによって好ましいコミュニティが異なることもあり，この実験のように複数の基準で抽出して比較することが大切であると言える．

5　結論

本研究では，隣接数に着目してハイパーグラフ上のコミュニティの 4 つの定義を与えた．これらは，グラフ上の adj-コミュニティを特殊ケースとして含んでいる．それぞれ，性質を満たすグラフを作成して最小カットアルゴリズムにより効率よくコミュニティが抽出できることを示した．さらに，共著データをハイパーグラフとして表したときに抽出されるこれら 4 つのコミュニティを比較検討した．この結果から，対象とするハイパーグラフの特徴，特にハイパーエッジの大きさによってどの定義によるコミュニティを用いるかを決定するべきであるということが分かった．

謝辞

本研究は科研費（22510135）および栢森情報科学振興財団の助成を受けたものである．

参考文献

[1] R. K. Ahuja, T. L. Magnanti and J. B. Orlin. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, 1993.

[2] M. J. Barber. Modularity and community detection in bipartite networks. Physical Review E, 76, 066102, 2007.

[3] M. Brinkmeier, S. Recknagel, and J. Werner. Communities in graphs and hypergraphs. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 869-872, 2007.

[4] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 150-160, 2000.

[5] W. Imrich, S. Klavžar, Product graphs: structure and recognition, John Wiley & Sons, New York, 2000.

[6] H. Ino, M. Kudo and A. Nakamura. Partitioning of web graphs by community topology. In Proceedings of the 14th international conference on World Wide Web, 661-669, 2005.

[7] Vladimir Batagelj and Andrej Mrvar. Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/ (2011/11/29 アクセス)

[8] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[9] 張明超, 高橋里司, 繁野麻衣子, 最大密度部分集合問題と近似２分探索による解法. 日本オペレーションズ・リサーチ学会和文論文誌, Vol. 53, 1-13, 2010.