ノードが複数の属性を持つグラフにおけるコミュニティ検出

伊藤　寛祥†　　駒水　孝裕††　　天笠　俊之††　　北川　博之††

† 筑波大学大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1 丁目 1 － 1
†† 筑波大学計算科学研究センター　〒 305-8573 茨城県つくば市天王台 1 丁目 1 － 1
E-mail: †hiro.3188@kde.cs.tsukuba.ac.jp, ††taka-coma@acm.org, {amagasa,kitagawa}@cs.tsukuba.ac.jp

あらまし　コミュニティ検出は，グラフデータの分析における重要な技術である．実世界に存在するグラフデータは多くの場合，ノードを特徴づける複数種類の属性を保持している．この属性を用いることで，コミュニティ検出に加え，そのコミュニティの性質を検出できると考えられる．また，属性値にも，類似性に基いたクラスタが存在すると考えられる．コミュニティと属性値のクラスタは相互に関連すると考えられ，また，これらを同時に検出することで，コミュニティ検出と属性値のクラスタの検出が相補的に行われ，精度が向上することが期待されるが，複数の属性のクラスタを考慮したコミュニティ検出手法は存在しない．そこで，本研究では，コミュニティ検出と属性値のクラスタの検出，コミュニティと属性値のクラスタの間の関連性を同時に検出する手法を提案する．実データを用いた，ベースライン手法との比較実験により，コミュニティ検出の精度が 11%，属性値のクラスタ検出の精度が 7∼22%程度向上することを確認した．

キーワード　コミュニティ検出，非負値行列分解

1. はじめに

コミュニティ検出とは，グラフにおいてノード間が密に接続しているノードの部分集合を発見する問題である．コミュニティに属するノードは同様の性質をもつことが多く，この特性から，コミュニティ検出はノードの性質の推定 [5], [7], [18] や，コミュニティへの情報の推薦 [8]，ノードの相互関係の分析 [2] など，さまざまなアプリケーションに応用されている．

コミュニティ検出をアプリケーションに応用する際には，コミュニティがどのような性質をもつかを検出することが重要であり，近年よく研究されている．ノードのテキスト情報を用いてコミュニティの性質を検出する手法として，グラフ中のコミュニティと，そのコミュニティがどのようなトピックに関心をもつかを検出する手法が多く提案されている [13], [14], [17], [19]．上に挙げた研究は，コミュニティとトピックが一対一対応というモデリングをしている．しかしながら，コミュニティは複数のトピックに関連があり，また，トピックは複数のコミュニティから関連付けられるという性質がある．学術研究を例にすると，ある研究者コミュニティは “データベース” と “データマイニング” の 2 つの研究分野に取り組むことがあり，また，“データベース” の研究分野は複数の研究者コミュニティから研究される．これに対し，Yin らはコミュニティとトピックを確率モデルで同時にモデリングし，その間の関連性を検出する手法を提案した [20]．この研究では，コミュニティは複数のトピックに関心を持ち，トピックは複数のコミュニティと関連するというモデリングを行っている．

ここで，実世界に存在するグラフデータの多くはノードが複数種類の属性を保持する．このようなグラフの例として，共著に基づく研究者ネットワークや，Twitter などのソーシャルメディアにおけるソーシャルネットワークがそれに該当する．研究者ネットワークにおいては，研究者をノードとすると，参加したことがある研究会議や，過去に投稿した論文，そこに含まれる単語などがノードの属性に対応する．Twitter においては，ユーザをノードとすると，ユーザが発信したツイートに含まれる単語やハッシュタグ，居住地などがノードの属性に対応する．このような属性は，ノードと属性値との関係における偏りから，属性値のクラスタが存在すると考えられ，また，属性値のクラスタは特定のコミュニティと関連すると考えられる．研究者ネットワークにおける属性のひとつである研究会議を例に挙げると，ある研究者コミュニティは特定の分野の研究を行うため，特定の分野の研究会議に参加し，また，その偏りから，研究会議にもクラスタが存在する．このような，ノードの属性値のクラスタの情報は，テキスト情報と同様に，コミュニティの性質や関心を推定する際に有用な情報になると考えられる．

これに対応し，ノードとエッジの種類が複数存在するグラフをクラスタリングする手法が提案された [12], [16]．これらの研究はいずれも，すべてのノードのクラスタ数が同じという仮定を置いている．しかしながら，各種属性はそれぞれ異なる粒度のクラスタをもち，コミュニティは複数のクラスタに関連をもつ．たとえば，データベースに関連する研究会議は，データベースシステムや，データマイニングといったトピックを取り扱うことから，単語のクラスタより研究会議のクラスタの方が粒度が大きいと考えられる．また，研究者コミュニティは多くの場合複数の研究テーマに取り組み，研究会議には複数の研究者コミュニティの研究者が出席する．

そこで，本研究では，ノードが複数の属性をもつグラフから，コミュニティ，属性値のクラスタ，コミュニティと属性値のクラスタとの関連性を検出する手法を提案する（図 1）．属性値のクラスタとは，ノードがもつ各種属性において，関連性の強い属性値の集合のことを指す．単語の属性を例にする

と，“database” と “query” など，よく共起する単語の集合が単語の属性値のクラスタとなる．会議の属性を例にすると，“SIGMOD” や “VLDB” など，分野が近い会議の集合が会議の属性値のクラスタとなる．コミュニティと属性クラスタ，その間の関連性を検出することで，コミュニティがどのような事柄に関心があるかを知ることができ，さらに，コミュニティ間がどのような事柄で共通しているかを検出することができる．

本研究では，非負値行列分解 (Non-negative Matrix Factorization，以降 NMF と表記する) に基づき，コミュニティ検出，属性値のクラスタ検出，コミュニティと属性値のクラスタ間の関係の検出を行う．ここで，これらをひとつの損失関数で表現することで，コミュニティの検出，属性値のクラスタの検出を連携させ，精度向上を狙う．すなわち，グラフ中のエッジ情報が欠損している場合でも，どの属性値のクラスタに関連するかに基づき，情報を補うことができ，また，属性値にノイズが存在する場合でも，どのコミュニティの属性値であるかを考慮することで，属性値のクラスタリングの精度が向上することが期待される．

実験では論文データベース DBLP [1] を用い，手法の有用性の検証を行った．実験より，本手法で，研究者コミュニティ，属性値のクラスタ，その間の関係性の検出できることを示した．また，既存の手法と比較して，コミュニティの検出精度が 11%，属性クラスタの検出精度が 7~22%程度向上することを示した．

本研究の貢献は以下である．

- コミュニティと複数種類の属性の属性クラスタ，その間の関連性の検出という新しい問題を定義し，その検出に成功した．
- 非負値行列分解に基づく，上記の問題に対応した手法を提案し，効率的な最適化アルゴリズムを導出した．
- 既存のコミュニティ検出手法，属性クラスタの検出手法と比較し，コミュニティの検出精度を約 11%，属性値のクラスタの検出精度を 7~22%向上させた．

本論文の構成は以下である．2. 節で本研究と関連する研究に関して述べる．3. 節で本研究で取り組む問題の定義を行い，4. 節で本論文の提案手法に関して述べる．5. 節で，本研究で行った実験に関して述べ，6. 節で結論を述べる．

2. 関 連 研 究

コミュニティ検出に関する手法はさまざまなアプローチが存在し，グラフの分割に基づく手法 [7], [15]，確率モデルに基づく手法 [21]，NMF に基づく手法 [9], [18] などが存在する．グラフの分割に基づく手法としては Newman-Girvan 法 [7]，n-cut [15] などが主要な手法として挙げられる．確率モデルに基づく手法としては Zhang らの SSN-LDA [21] が挙げられ，LDA [3] に基づきグラフ中のコミュニティを潜在確率変数としてモデリングし，ノードの多項分布としてコミュニティを検出した．NMF に基づく手法として，Gegick らはグラフの隣接行列に対して NMF を適用することで，グラフ中のコミュニティを検出する Symmetric-NMF [9] を提案した．Yang らは，オーバーラップしたコミュニティを高速に検出する NMF に基づく手法

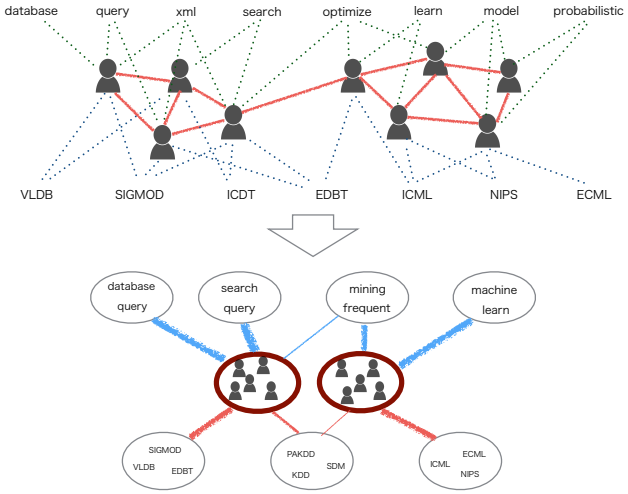


図 1　**本研究の目的**．本研究では，ノードが複数種類の属性を持つグラフから，コミュニティ，属性値のクラスタ，その間の関連性を検出することを目的とする．

BigCLAM [18] を提案した．これらの手法は，グラフ中のノードの属性を考慮せず，グラフのエッジ情報のみに基づいたコミュニティ検出手法である．このような研究は，[4] によくまとめられているため，そちらも参照されたい．

近年では，コミュニティの検出に加え，コミュニティがもつ性質を検出する手法が提案されている．Yang らは，コミュニティの検出に対し，ユーザが生成した情報を付与する手法 CESNA [19] を提案した．Wang らは，NMF に基づき，グラフにおけるコミュニティと，そのコミュニティに属するノードの属性のクラスタを同時に検出する手法 SCI [17] を提案した．Michael らは，トピックモデルに基づきコミュニティの検出と，コミュニティのトピックを検出する手法 Author-Topic model [14] を提案した．本研究はテキストデータのみなど，ひとつの属性ではなく，複数種類の属性を用いる点でこれらの研究と異なる．

ノードとエッジの種類が複数種類存在するグラフデータに対するクラスタリング手法も近年よく研究されている．Yin らは，NMF に基づいて，複数種類存在するノードをクラスタリングする手法を提案した [12]．Sun らは，ランキングが上位のノードはクラスタにおける影響力も大きいという観点から，ノードのクラスタリングとノードのランキングを同時に行う手法を提案した [16]．これらの手法は各種ノードを単一のクラスタにクラスタリングする手法であるが，本研究の手法は，各種ノードをそれぞれ別種類とみなしてクラスタリングし，かつ，コミュニティと属性クラスタとの関係性を明示的に得られるという点で，これらの研究と異なる．

本研究の立ち位置は，表 1 にまとめられる．本研究は，ノードがもつ複数種類の属性を考慮し，コミュニティと属性値のクラスタ，その間の関連性を検出する初の手法である．

3. 問 題 定 義

本研究では，複数種類の属性をもつグラフデータを対象に分

表 1　コミュニティ検出に関連する研究との比較

コミュニティと属性値のクラスタとの関連性	エッジ情報のみ	テキスト情報を利用	複数種類の属性情報を利用
得られない	[7], [9], [15], [18], [21]	[14], [17], [19]	[12], [16]
得られる	–	[20]	本研究

析を行い，コミュニティ，属性値のクラスタ，およびコミュニティと属性値のクラスタとの関連性を検出する．本節では，提案手法の入力である，ノードが複数種類の属性をもつグラフデータの定義，および，本研究の目的である，コミュニティ検出，属性値のクラスタ検出，その間の関連性の検出のそれぞれの問題定義を与える．

ノードの集合とそれら間の関係性は，以下で定義される重み付きグラフで表現される．

定義 3.1 (重み付きグラフ)　$\mathcal{G}' = \langle \mathcal{V}, \mathcal{E}, \mathcal{W} \rangle$ を重み付きグラフとし，$\mathcal{V}$ をグラフ中のノード集合，$\mathcal{E}$ をグラフ中のエッジ集合，$\mathcal{W} : \mathcal{E} \to \mathbb{R}^+$ をエッジの重みとし，$e \in \mathcal{E}$ から実数 $w \in \mathbb{R}^+$ への写像とする．□

さらに，グラフ中の各ノードは複数の属性で特徴づけられる．ここで，本研究では属性値は数値ではなく，集合で与えられるものとする．ノードと属性値との間の関係性は二部グラフで表現される．これを属性グラフとし，以下に定義する．

定義 3.2 (属性グラフ)　属性 $t \in \mathcal{T}$ の**属性値**集合を $\mathcal{X}_t$ とし，ノードと $\mathcal{X}_t$ との 2 部グラフを属性グラフ $\mathcal{G}_t = \langle \mathcal{V} \cup \mathcal{X}_t, \mathcal{E}_t, \mathcal{W}_t \rangle$ とする．$\mathcal{E}_t$ はエッジ集合であり，$\mathcal{W}_t : \mathcal{E}_t \to \mathbb{R}^+$ はエッジの重みで，エッジ $e \in \mathcal{E}_t$ から実数 $w \in \mathbb{R}^+$ への写像とする．□

本研究では，ノードは複数の属性 $\mathcal{T}$ で特徴づけられるものとする．ノードが複数種類の属性をもつグラフの定義は以下である．

定義 3.3 (ノードが複数種類の属性をもつグラフ)　重み付きグラフ $\mathcal{G}' = \langle \mathcal{V}, \mathcal{E}, \mathcal{W} \rangle$ と，複数の属性グラフ $\{\mathcal{G}_t\}_{t \in \mathcal{T}}$（ここで，$\mathcal{G}_t = \langle \mathcal{V} \cup \mathcal{X}_t, \mathcal{E}_t, \mathcal{W}_t \rangle$ である）が与えられたとき，これらによって構成されるグラフ $\mathcal{G} = \left\langle \mathcal{G}', \{\mathcal{G}_t\}_{t=1}^{T} \right\rangle$ を**ノードが複数種類の属性をもつグラフ**とする．□

本研究では，ノードが複数種類の属性をもつグラフから，コミュニティ，属性値のクラスタ，コミュニティと属性値のクラスタとの関連性の 3 つを検出することを目的とする．はじめに，それぞれの問題に関する本研究での仮定を述べる．

a)　コミュニティ

ノードが複数種類の属性をもつグラフ $\mathcal{G}$ において，以下の 2 つの性質をもつノード集合を**コミュニティ**とする．(1) コミュニティに属するノード間はコミュニティ外と比較して密にエッジが接続されている．(2) コミュニティに属するノードは似た性質をもつ．ここで，本研究では，各ノードは複数のコミュニティに属するものとする．具体的には，コミュニティ数 l が与えられた時，コミュニティ $c \in \mathcal{C}$ に属するノード $n \in \mathcal{V}$ は確率分布 $p(n \mid c)$ で与えられるものとする ($|\mathcal{C}| = l$)．

b)　属性値のクラスタ

ノードが複数種類の属性をもつグラフ $\mathcal{G}$ が与えられた時，属性 $t \in \mathcal{T}$ において，属性値同士がよく関連している属性値の集合を属性 $t \in \mathcal{T}$ の**属性値のクラスタ**とする．ここで，本研究では，各属性値は複数のクラスタに属するものとする．具体的には，属性 $t \in \mathcal{T}$ の属性値のクラスタ数 $k^{(t)}$ が与えられた時，属性値のクラスタ $s_t \in \mathcal{S}_t$ に属する属性値 x は確率分布 $p(x \mid s_t)$ で与えられるものとする ($|\mathcal{S}_t| = k^{(t)}$)．

c)　コミュニティと属性値のクラスタとの関連性

コミュニティ c と属性 $t \in \mathcal{T}$ の属性値のクラスタ s_t が与えられた時，c が s_t と関連する確率を**コミュニティと属性値のクラスタとの関連性**とする．ここで，コミュニティは複数の属性値のクラスタに関連するものとする．具体的には，コミュニティ c と属性 $t \in \mathcal{T}$ の属性値のクラスタの集合 $\mathcal{S}_t$ が与えられた時，コミュニティ c と属性 $t \in \mathcal{T}$ のクラスタ $s_t \in \mathcal{S}_t$ との関連性は確率分布 $p(s_t \mid c)$ で与えられるものとする．

以上を踏まえ，本研究の目的である，コミュニティ検出，属性値のクラスタ検出，コミュニティと属性値のクラスタとの関連性の検出を以下に定義する．

定義 3.4 (本研究の目的)　ノードが複数種類の属性をもつグラフ $\mathcal{G}$ が与えられた時，以下の 3 つを検出することである．

- コミュニティ：
  $\forall n \in \mathcal{V},\ \forall c \in \mathcal{C},\ p(n \mid c)$　(ここで，$|\mathcal{C}| = l$)
- 属性値のクラスタ：
  $\forall t \in \mathcal{T},\ \forall x \in \mathcal{X}_t,\ \forall s_t \in \mathcal{S}_t,\ p(x \mid s_t)$　(ここで，$|\mathcal{S}_t| = k^{(t)}$)
- コミュニティと属性値のクラスタとの関連性：
  $\forall c \in \mathcal{C},\ \forall t \in \mathcal{T},\ \forall s_t \in \mathcal{S}_t,\ p(s_t \mid c)$　□

4. 提 案 手 法

本節では，NMF にもとづき，複数種類の属性をもつグラフから，コミュニティ，属性値のクラスタ，その間の関連性を検出する手法を提案する．具体的には，コミュニティと属性値のクラスタを NMF に基づきモデリングし，コミュニティと属性値のクラスタを関連付ける行列を新たに設定し，それらを同時に行列分解することで，3. 節で挙げた情報を検出することを目指す．

4.1 損 失 関 数

ここでは，本研究で設定する損失関数に関して記述する．はじめに，入力となる行列の構成，コミュニティ検出，属性値のクラスタの検出，その間の関連性の検出に関する損失関数の設定に関して記述する．つづいて，それらを統合した損失関数に関して詳細を記述する．

入力行列の構成: 本手法では，ノードが複数種類の属性を持つグラフを，グラフの隣接行列 $A \in \mathbb{R}^{|V| \times |V|}$ と，ノードと 属性

$t \in \mathcal{T}$ の属性値との関係を表現する行列 $X^{(t)} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{X}^{(t)}|}$ で表現する．行列 A の各要素は $A_{u,v} = w_{u,v} / \sum_{u,v} w_{u,v}$ とする．ここで，$w_{u,v} = \mathcal{W}(e_{u,v})$ はノード u とノード v の間のエッジの重みを指し，$A_{u,v}$ は複数種類の属性をもつグラフ $\mathcal{G}$ においてノード u,v が接続される確率 $p(u,v)$ を表す．行列 $X^{(t)}$ の各要素は $X^{(t)}_{u,x} = w^{(t)}_{u,x} / \sum_{u,x} w^{(t)}_{u,x}$ とする．ここで，$w^{(t)}_{u,x} = \mathcal{W}_t(e^{(t)}_{u,x})$ はノード u と属性 t における属性値 x とのエッジの重みを指し，$X^{(t)}_{u,x}$ は複数種類の属性をもつグラフ $\mathcal{G}$ においてノード u が属性値 x をもつ確率 $p(u,x)$ を表す．

コミュニティ検出の損失関数：グラフ中のコミュニティを行列 $U^* \in \mathbb{R}^{|\mathcal{V}| \times l}$ で表現し，各要素 $U^*_{u,c}$ をノード u がコミュニティ c に属する確率 $p(u \mid c)$ とする．ここで l はコミュニティ数である．ノード u と v がコミュニティ c においてエッジが存在する確率は $U^*_{u,c}U^*_{v,c}$ で表現され，複数種類の属性をもつグラフにおいてノード u と v でエッジが存在する確率 $p(u,v)$ は $\sum_{c \in \mathcal{C}} U^*_{u,c}U^*_{v,c}$ で表現される．すなわち，グラフ中のエッジの構造を最もよく表現する U^* は以下の損失関数を最小化する U^* である．

$$\underset{U^* \geqq 0}{\arg\min} \left\| A - U^*(U^*)^T \right\|_F^2 \tag{1}$$
$$s.t.\ \ \forall 1 \leqq p \leqq l,\ \ \left\| U^*_{\cdot,p} \right\|_1 = 1$$

属性値のクラスタの検出に関する損失関数：属性 t の属性値のクラスタを表す行列を $V^{(t)} \in \mathbb{R}^{|\mathcal{X}^{(t)}| \times k^{(t)}}$ で表現し，各要素 $V^{(t)}_{x,s_t}$ は属性値 x がクラスタ s_t に属する確率 $p(x \mid s_t)$ とする．ここで $k^{(t)}$ は属性 t のクラスタ数である．本研究では $V^{(t)}$ を，グラフのノードと属性値との関係から検出する．グラフのノードと属性値のクラスタとの関連を行列 $U^{(t)} \in \mathbb{R}^{|\mathcal{V}| \times k^{(t)}}$ で表現し，各要素 $U^{(t)}_{u,s_t}$ は，ノード u が属性値のクラスタ s_t に属する確率 $p(u \mid s_t)$ とする．ノード u が属性 t の属性値のクラスタ s_t において属性値 x に関連する確率 $p(u,c \mid s_t)$ は $U^{(t)}_{u,s_t}V^{(t)}_{x,s_t}$ で表現され，グラフにおいてノード u と属性値 x にエッジが存在する確率 $p(u,x)$ は $\sum_{s_t \in \mathcal{S}_t} U^{(t)}_{u,s_t}V^{(t)}_{x,s_t}$ で表現される．すなわち，グラフ中のエッジの構造を最もよく表現する行列 $U^{(t)}$，$V^{(t)}$ は以下の損失関数を最小化する行列である．

$$\underset{U^{(t)},V^{(t)} \geqq 0}{\arg\min} \left\| X^{(t)} - U^{(t)}(V^{(t)})^T \right\|_F^2 \tag{2}$$
$$s.t.\ \ \forall 1 \leqq r \leqq k^{(t)}, \left\| U^{(t)}_{\cdot,r} \right\|_1 = 1, \left\| V^{(t)}_{\cdot,r} \right\|_1 = 1$$

コミュニティと属性値のクラスタの関連性の検出に関する損失関数：本研究では，ノードのコミュニティと属性値のクラスタ t との間の関連を示す行列を $R^{(t)} \in \mathbb{R}^{l \times k^{(t)}}$ で表現し，各要素 $R^{(t)}_{s_t,c}$ をコミュニティ c が属性値のクラスタ s_t に関連する確率 $p(s_t \mid c)$ とする．本研究では，ノードがどのコミュニティに属するかを表現する行列 U^* に対して $R^{(t)}$ で線形写像することで，ノードがどの属性値のクラスタに属するかを表現する行列 $U^{(t)}$ に変換することを考える．すなわち，以下の損失関数を最小化するような $R^{(t)}$ を求めることで，コミュニティと属性値のクラスタとの関係を得ることができる．

$$\underset{U^{(t)},U^*,R^{(t)} \geqq 0}{\arg\min} \left\| U^{(t)} - U^* R^{(t)} \right\|_F^2 \tag{3}$$
$$s.t.\ \ \forall 1 \leqq r \leqq k^{(t)},\ \forall 0 \leqq p \leqq l,$$
$$\left\| U^{(t)}_{\cdot,r} \right\|_1 = 1, \left\| U^*_{\cdot,p} \right\|_1 = 1, \left\| R^{(t)}_{p,\cdot} \right\|_1 = 1$$

また，この項は，ノードと属性値のクラスタとの関連を表す行列を，ノードがどのコミュニティに属するかを表現する行列と，コミュニティと属性値のクラスタの行列に分解する，非負値行列分解と解釈することができる．すなわち，この項を最適化することで，著者がどの属性値のクラスタに関連するかという情報を，グラフ構造に基づくコミュニティの検出に対して付与することを意味している．

統合した損失関数：本手法では，コミュニティ検出，属性値のクラスタの検出，その間の関連性をひとつの損失関数でモデリングする．こうすることで，コミュニティ検出，属性値のクラスタの検出がそれぞれの情報を補い合うことを狙いとする．コミュニティ検出に関する損失関数，属性値のクラスタの検出に関する損失関数，コミュニティと属性値のクラスタの関連性の検出に関する損失関数をすべて合わせると，以下の損失関数が得られる．

$$L = \underset{U^*,\ \{U^{(t)},\ V^{(t)},\ R^{(t)}\}_{t=1}^T \geqq 0}{\arg\min} \left\| A - U^*(U^*)^T \right\|_F^2 + \sum_{t=1}^{T} \left\{ \left\| X^{(t)} - U^{(t)}(V^{(t)})^T \right\|_F^2 + \lambda_t \left\| U^{(t)} - U^* R^{(t)} \right\|_F^2 \right\} \tag{4}$$
$$s.t.\ \ \forall 1 \leqq r \leqq k^{(t)},\ \forall 1 \leqq p \leqq l,\ \forall 1 \leqq t \leqq T,$$
$$\left\| U^*_{\cdot,p} \right\|_1 = 1, \left\| U^{(t)}_{\cdot,r} \right\|_1 = 1, \left\| V^{(t)}_{\cdot,r} \right\|_1 = 1, \left\| R^{(t)}_{p,\cdot} \right\|_1 = 1$$

ここで λ_t はユーザが与えるパラメータであり，属性 i の属性値のクラスタがコミュニティの検出に対して与える影響力の大きさとなる．この λ_t が大きくなるほど，属性値のクラスタがコミュニティの検出に対して与える影響力が大きくなる．

4.2 最 適 化

式 4 はすべての変数に関して同時に凸にならない．本研究では [11] に基づいて各変数に関して更新式を導出し，それを繰り返し実行することで最適化を行う．KKT 条件は以下で表される．

$$U^* \geqq \mathbf{0},\ U^{(t)} \geqq \mathbf{0},\ V^{(t)} \geqq \mathbf{0},\ R^{(t)} \geqq \mathbf{0} \tag{5}$$
$$\nabla_{U^*} L \geqq \mathbf{0},\ \nabla_{U^{(t)}} L \geqq \mathbf{0},\ \nabla_{V^{(t)}} L \geqq \mathbf{0},\ \nabla_{R^{(t)}} L \geqq \mathbf{0} \tag{6}$$
$$U^* \odot \nabla_{U^*} L = \mathbf{0},\ U^{(t)} \odot \nabla_{U^{(t)}} L = \mathbf{0},$$
$$V^{(t)} \odot \nabla_{V^{(t)}} L = \mathbf{0},\ R^{(t)} \odot \nabla_{R^{(t)}} L = \mathbf{0} \tag{7}$$

続いて各変数に関する勾配は，以下で表される．

$$\nabla_{U^*} L = -2A^T RU^* + 2U^*(U^*)^T U^* + \sum_{t=1}^{T} \lambda_t(-U^{(t)}(R^{(t)})^T + U^* R^{(t)}(R^{(t)})^T) \quad (8)$$

$$\nabla_{U^{(t)}} L = -X^{(t)}V^{(t)} + U^{(t)}(V^{(t)})^T V^{(t)} + \lambda_t(U^{(t)} - U^* R^{(t)}) \quad (9)$$

$$\nabla_{V^{(t)}} L = -(X^{(t)})^T U^{(t)} + (U^{(t)})^T U^{(t)}(V^{(t)})^T \quad (10)$$

$$\nabla_{R_i} L = -(U^*)^T U^{(t)} + (U^*)^T U^* R^{(t)} \quad (11)$$

式 7 に対して勾配を代入することで，各変数に関して更新式を導出した．

$$U^* \leftarrow U^* \odot \frac{A^T U^* + \sum_{i=1}^{T} \lambda_t U^{(t)}(R^{(t)})^T}{2U^*(U^*)^T U^* + \sum_{i=1}^{T} U^* R^{(t)}(R^{(t)})^T} \quad (12)$$

$$U^{(t)} \leftarrow U^{(t)} \odot \frac{X^{(t)}V^{(t)} + \lambda_t U^* R^{(t)}}{U^{(t)}(V^{(t)})^T V^{(t)} + \lambda_t U^{(t)}} \quad (13)$$

$$V^{(t)} \leftarrow V^{(t)} \odot \frac{(X^{(t)})^T U^{(t)}}{(U^{(t)})^T U^{(t)}(V^{(t)})^T} \quad (14)$$

$$R^{(t)} \leftarrow R^{(t)} \odot \frac{U^* U^{(t)}}{(U^*)^T U^* R^{(t)}} \quad (15)$$

式 12，13，14，15 を繰り返し適用することで，損失関数は単調減少し，ある程度収束したところで行列分解が完了する．

式 4 における条件，$\left\|U^*_{:,p}\right\|_1 = 1$，$\left\|U^{(t)}_{:,r}\right\|_1 = 1$，$\left\|V^{(t)}_{:,r}\right\|_1 = 1$，$\left\|R^{(t)}_{p,:}\right\|_1 = 1$ を満たすため，以下に示す正規化の式を，最適化の際に適用する．

$$U^* \leftarrow U^* \left(Q^*\right)^{-1} \quad (16)$$

$$U^{(t)} \leftarrow U^{(t)} \left(Q^{(t)}\right)^{-1} \quad (17)$$

$$V^{(t)} \leftarrow V^{(t)} Q^{(t)} \quad (18)$$

$$R^{(t)} \leftarrow R^{(t)} \left(Q^{R^{(t)}}\right)^{-1} \quad (19)$$

ここで，Q^*, $Q^{(t)}$, Q^R は

$$Q^* = Diag\left(\sum_{i=1}^{|\mathcal{V}|} U^*_{i,1}, \sum_{i=1}^{|\mathcal{V}|} U^*_{i,2}, \cdots, \sum_{i=1}^{|\mathcal{V}|} U^*_{i,l}\right) \quad (20)$$

$$Q^{(t)} = Diag\left(\sum_{i=1}^{|\mathcal{X}^{(t)}|} U^{(t)}_{i,1}, \sum_{i=1}^{|\mathcal{X}^{(t)}|} U^{(t)}_{i,2}, \cdots, \sum_{i=1}^{|\mathcal{X}^{(t)}|} U^{(t)}_{i,k^{(t)}}\right) \quad (21)$$

$$Q^{R^{(t)}} = Diag\left(\sum_{i=1}^{k^{(t)}} R^{(t)}_{i,1}, \sum_{i=1}^{k^{(t)}} R^{(t)}_{i,2}, \cdots, \sum_{i=1}^{k^{(t)}} R^{(t)}_{i,l}\right) \quad (22)$$

である．

以上をまとめると，最適化のアルゴリズムは Algorithm1 で示される．ここで，行列の正規化は次の行列の更新の前に行う．これは，行列を更新した際，各列の和が 1 になることを保証するためである．行列の正規化を行わない状態では，各行列の要素の値域が異なる場合が想定され，行列を更新する際に値域が大きい行列の影響が大きくなる可能性がある．このため，本研究では，Algorithm1 で示した手順で最適化を行う．

Algorithm 1 Optimization Algorithm

Require: $A, \{X^{(t)}\}_{t\in\mathcal{T}}, \{\lambda_t\}_{t\in\mathcal{T}}, \epsilon$
Ensure: $U^*, \{U^{(t)}, V^{(t)}, R^{(t)}\}_{t\in\mathcal{T}}$
$U^*, \{U^{(t)}, V^{(t)}, R^{(t)}\}_{t\in\mathcal{T}} \leftarrow$ random non-negative init
$\epsilon' \leftarrow maxInt$, $\epsilon \leftarrow \frac{\epsilon'}{2}$
while $abs(\epsilon' - \epsilon) \geqq \epsilon$ **do**
  $U^* \leftarrow U^* \odot \frac{A^T U^* + \sum_{t\in\mathcal{T}} \lambda_t U^{(t)}(R^{(t)})^T}{2U^*(U^*)^T U^* + \sum_{t\in\mathcal{T}} U^* R^{(t)}(R^{(t)})^T}$
  $U^* \leftarrow U^*(Q^*)^{-1}$
  for $t \in \mathcal{T}$ **do**
    $U^{(t)} \leftarrow U^{(t)} \odot \frac{X^{(t)}V^{(t)} + \lambda_t U^* R^{(t)}}{U^{(t)}(V^{(t)})^T V^{(t)} + \lambda_t U^{(t)}}$
    $V^{(t)} \leftarrow V^{(t)} \odot \frac{(X^{(t)})^T U^{(t)}}{(U^{(t)})^T U^{(t)}(V^{(t)})^T}$
    $U^{(t)} \leftarrow U^{(t)} \left(Q^{(t)}\right)^{-1}$
    $V^{(t)} \leftarrow V^{(t)} Q^{(t)}$
    $R^{(t)} \leftarrow R^{(t)} \odot \frac{U^* U^{(t)}}{(U^*)^T U^* R^{(t)}}$
    $R^{(t)} \leftarrow R^{(t)} \left(Q^{R^{(t)}}\right)^{-1}$
  end for
  $\epsilon' \leftarrow \epsilon$
  $\epsilon \leftarrow L\left(U^*, \{U^{(t)}, V^{(t)}, R^{(t)}\}_{t\in\mathcal{T}}\right)$
end while

5. 実　　験

本節では，提案手法の有用性に関して議論する．5.1 節で，実験に使用したデータセットと提案手法の適用方法に関して述べる．5.2 節で，実際に検出されたコミュニティ，属性値のクラスタ，コミュニティと属性値のクラスタとの関係を掲載する．5.3 節で，コミュニティ検出，属性値のクラスタ検出の精度に関して，本手法と関連する手法とを比較する．最後に，5.4 節で，パラメータの変化で，出力にどのような影響を与えるか議論する．

5.1 データセット

本研究では論文データベース DBLP を用い，手法の検証，評価を行った．Digital Bibliography Project (DBLP) [1] はコンピュータサイエンス分野の論文データベースである．本研究では，4 つの研究分野 (データベース，データマイニング，情報検索，機械学習) に [6] で人手で分類された著者 4019 名が，人手で 4 つの分野に分類された特定の 20 の会議 (1 分野あたり 5 つの会議) で 2009 年までに発表した論文 10491 件をデータセットとして用いる．

本研究では論文の著者をグラフのノードとし，共著関係をグラフのエッジ，共著で論文を発表した回数をグラフの重みとする．また，ノードがもつ属性としては，“単語” と “論文”，“会議” とする．各属性に関して以下に詳細を記載する．

- **単語：** 単語の集合を，この属性の属性値集合とする．著者がこれまでに発表した論文に含まれる単語に対してエッジが存在するものとし，エッジの重みは著者がこれまで発表した論文に含まれる各単語の出現回数とする．ここで，本研究ではストップワードの除去と，ステミング処理を行った．属性値として使用する単語数は 1367 とした．
- **論文：** 論文の集合を，この属性の属性値集合とする．著者がこれまでに発表した論文に対してエッジが存在するものとし，エッジの重みは常に 1 とする．属性値として使用する論文数は上記の項目で述べたとおりである．
- **会議：** 会議の集合を，この属性の属性値集合とする．著者が発表した論文がこれまでに掲載されたことのある会議に対してエッジが存在するものとし，エッジの重みは著者が発表した論文が各会議に掲載された回数とする．属性値として使用する会議数は上記の項目で述べたとおりである．

5.2 実際に検出された出力

本節では，本手法をデータセットに対して適用し，検出されたコミュニティと属性値のクラスタ，その間の関係性に関して議論する．ここでは，コミュニティ数と単語の属性値のクラスタ数を 50，会議のクラスタ数を 4 に設定して実験を行った．図 2(a)， 図 2(b)， 図 2(c) は，それぞれ本手法で検出された出力である．図中の赤い四角形がコミュニティを表し，中に列挙されている名前が，コミュニティに属するメンバーである．名前の脇の数値は，研究者 u が研究コミュニティ c に属する確率 $p(u \mid c) = U^*_{u,c}$ を表す．ここではコミュニティに属する確率が 0.01 以上の研究者を記載している．図中の灰色の四角形が研究会議の属性値のクラスタ，青色の四角形が単語の属性値のクラスタを表現しており，四角形の中に，クラスタを構成する属性値が記述されている．属性値の脇の数値は，属性 t の，ある属性値 x が属性値のクラスタ s_t に属する確率 $p(x \mid s_t) = V^{(t)}_{x,s_t}$ を表し，確率の大きい属性値から上位 5 つを記載している．コミュニティと属性値のクラスタを結ぶ線は，コミュニティと属性値のクラスタが関連していることを表しており，線の脇の数値はコミュニティ c が，属性 t の，ある属性値のクラスタ s_t と関連する確率 $p(s_t \mid c) = R^{(t)}_{s_t,c}$ を表している．また，ここでは研究コミュニティと関連度が大きい上位 3 つの属性値のクラスタを記載している．

図 2(a) は本手法で検出された，“*jiawei han*” が所属する研究コミュニティと，それに関連する属性値のクラスタを記載したものである．図 2(a) を見ると，“*jiawei han*” は多くの中国人の研究者と研究コミュニティを形成し，“*KDD,ICDM,SDM,PAKDD,VLDB*” というデータマイニングに関係する会議によく参加し，“*cluster*”，“*graph classif*” といった研究テーマに関連しているという結果が出力されたことがわかる．図 2(b) は本手法で検出された，“*michael stonebraker*” が所属する研究コミュニティと，それに関連する属性値のクラスタを記述したものである．図 2(b) を見ると，このコミュニティは “*SIGMOD,VLDB,PODS,EDBT,ICDT*” というデータベースに関係する会議によく参加し，“*system update*”,“*algorithm*”,“*query*” といった研究テーマに関連しているという結果が出力されたことがわかる．図 2(c) は本手法で検出された，“*michael i. jordan*” が所属する研究コミュニティと，それに関連する属性値のクラスタを記述したものである．図 2(c) を見ると，このコミュニティは “*NIPS,ICML,UAI,COLT,ECML*” という機械学習に関係する会議によく参加し，“*learn, model, network*”, “*expert model*”, “*predict process*” という研究テーマに関連するという結果が出力されたことがわかる．

5.3 関連手法との検出精度の比較

本節では，コミュニティの検出精度，属性値のクラスタの検出精度に関して，関連手法と比較を行う．本研究では，以下の手法と比較を行った．

- **NMF**: 通常の NMF．著者-単語 (A-T)，著者-論文 (A-P)，著者-会議 (A-C)，単語-論文 (T-P)，単語-会議 (T-C) の関係で作成される行列に対して，それぞれ NMF を適用する．(注1)
- **LCTA** [20]: コミュニティとトピックを同時にモデリングした確率的生成モデル．コミュニティとトピック，その間の関係性を検出することができる．
- **HINMF** [12]: 複数種類の属性をもつデータを対象とし，属性とデータとを同時にクラスタリングする NMF に基づく手法．

本研究では *Accuracy* で定量的に精度の評価を行う．Accuracy の定義は以下である．

定義 5.1 (Accuracy) ある集合中の要素 $n \in \mathcal{S}$，その要素のラベル S_n，ある手法から得られたラベル r_n が与えられた時，*Accuracy* は以下で定義される．

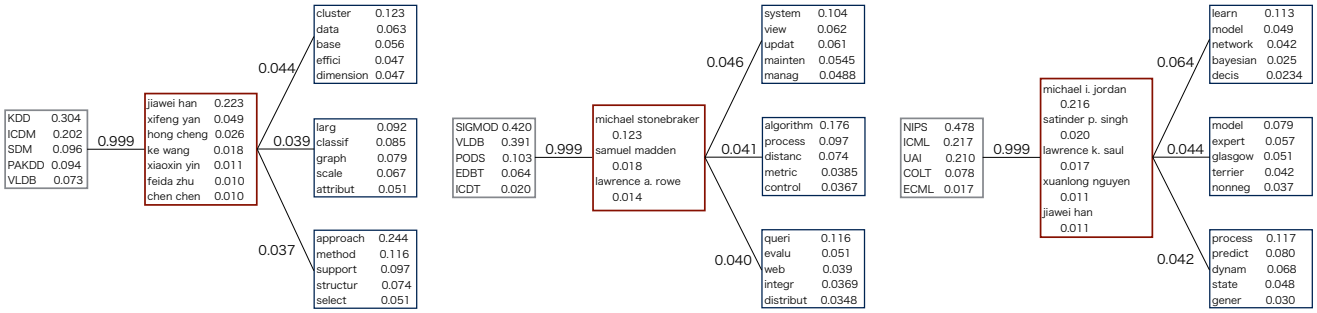
$$Accuracy = \frac{\sum_{n \in \mathcal{S}} \delta(s_n, map(r_n))}{|\mathcal{S}|} \qquad (23)$$

ここで，$|\mathcal{S}|$ 集合 $\mathcal{S}$ 全体の大きさ，$\delta(x,y)$ はデルタ関数で，$x = y$ のときに *1*，それ以外のときに *0* をとる関数である．$map(r_n)$ は著者 n が対応するラベル r_n をデータセット中のラベルにマッピングする関数である．最良のマッピングは *Kuhn-Munkres* アルゴリズム [10] によって発見する．□

ここでは，コミュニティ数を 4，各属性のクラスタ数を 4 にそれぞれ設定し，本手法と関連手法との *Accuracy* の比較を行った．正解セットは，“著者” と “研究会議” に関しては人手によるラベル，“論文” に関しては人手によるラベリングされた会議群から発表された論文，“単語” は各会議群から発表された論文に含まれる単語のうち最も出現する単語上位 10 件とし，Precision@10　で評価する．

表 2 は各手法で検出された各クラスタに関する *Accuracy* の値をまとめたものである．それぞれの手法で検出できないタスクは N/A となっている．表 2 をみると，コミュニティの検出精度 (Author) に関しては，本手法が関連研究と比較して 11% *Accuracy* が改善していることがわかる．また，属性値のクラスタの検出精度に関しては，Paper（論文）のクラスタ検

(注1)：NMF は共起関係に基づいたクラスタリング手法であるが，論文-会議は 1 対 1 の関係であるため，共起関係を得られず，クラスタリングを行うことができない．このため，ここでは論文-会議 (P-C) の行列は除外する．

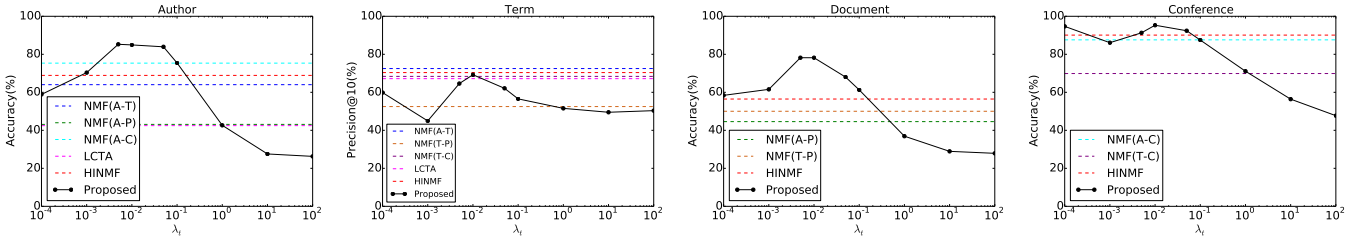


(a) コミュニティ; “*Jiawei Han*” (b) コミュニティ; “*Michael Stonebraker*” (a) コミュニティ; “*Michael I. Jordan*”

図 2 研究者コミュニティのメンバーと，コミュニティに関連する属性値のクラスタ

表 2 コミュニティ検出，属性値のクラスタ検出の *Accuracy*，*Precision*@10

Algorithm	Author	Term	Paper	Conference
NMF (A-T)	64.02 ± 5.73	**72.50 ± 0.00**	N/A	N/A
NMF (A-P)	43.12 ± 5.17	N/A	44.58 ± 5.89	N/A
NMF (A-C)	75.35 ± 6.85	N/A	N/A	87.60 ± 1.73
NMF (T-P)	N/A	52.50 ± 8.91	50.02 ± 7.93	N/A
NMF (T-C)	N/A	68.38 ± 5.88	N/A	69.88 ± 6.68
LCTA	42.47 ± 4.77	67.19 ± 5.65	N/A	N/A
HINMF	68.90 ± 9.08	70.38 ± 4.05	56.46 ± 3.08	90.10 ± 12.63
Proposed	**86.34 ± 2.39**	69.10 ± 6.34	**78.19 ± 9.87**	**97.20 ± 5.21**



(a) *Accuracy*; Author (b) *Accuracy*; Term (a) *Accuracy*; Paper (b) *Accuracy*; Conference

図 3 パラメータ λ_t を変化させた際の *Accuracy* の変化

出に関して 22%，Conference（会議）のクラスタ検出に関して 7% *Accuracy* が改善したことが分かる．これは，コミュニティ検出と属性値のクラスタの検出を同時行い，その間の関係性を同時にモデリングすることで，それぞれの検出精度が向上したことを示している．

5.4 パラメータ設定

本節では，提案手法に存在するパラメータ λ_t を変動させたときに，出力結果がどのように変化するかを議論する．本手法では，属性 t の属性値のクラスタがコミュニティの形成にどれだけの影響力を与えるかを，パラメータ λ_t で調節する．λ_t が大きいほど，属性 t の属性値のクラスタがコミュニティ形成に与える影響力が大きくなり，小さいほどコミュニティ形成に与える影響力が小さくなる．このため，パラメータの値が小さすぎると，属性値のクラスタがノード間のエッジの情報を補い，精度を向上させる効果が得られないと考えられる．また逆に，パラメータの値が大きすぎると，属性値のクラスタがコミュニティの形成に与える影響力が大きすぎて，ノード間のエッジの情報を無視する結果になり，精度が低下すると考えられる．

図 3(a), 図 3(b), 図 3(c), 図 3(d) はパラメータ λ_t を変動させた際の精度の変動をプロットしたものである．本手法の精度に加え，比較手法の精度を同時にプロットしている．本研究では，各パラメータをすべて同じ値に設定して実験を行った．図を見ると，どのタスクにおいても $\lambda_t = 0.005, 0.01$ のときに最も高い精度でクラスタリングが行われることが分かる．

6. 結　　論

本研究では，ノードがもつ複数種類の属性に着目し，グラフ中のコミュニティと属性のクラスタ，その間の関係性を検出する手法を提案した．実験より，本手法によって，コミュニティと属性のクラスタ，その間の関係性が検出できることを示し，また，既存手法と比較して，コミュニティと属性値のクラスタの検出精度が向上したことを示した．

今後の展望としては，コミュニティ，属性値のクラスタの時間的な変遷を検出する手法への拡張や，情報の伝播を推定する手法への拡張などが考えられる．

謝　　辞

本研究の一部は，JSPS 科研費 JP25240014，NICT 高度通信・放送研究開発委託研究「欧州との連携による公共ビッグデータの利活用基盤に関する研究開発」の助成を受けたものです．

文　　献

[1] DBLP. http://www.informatik.uni-trier.de/~ley/db/.

[2] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, Vol. 9, No. Sep, pp. 1981–2014, 2008.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[4] Santo Fortunato. Community detection in graphs. *Physics reports*, Vol. 486, No. 3, pp. 75–174, 2010.

[5] Mario Frank, Andreas P Streich, David Basin, and Joachim M Buhmann. Multi-assignment clustering for boolean data. *Journal of Machine Learning Research*, Vol. 13, No. Feb, pp. 459–489, 2012.

[6] Jing Gao, Feng Liang, Wei Fan, Yizhou Sun, and Jiawei Han. Graph-based consensus maximization among multiple supervised and unsupervised models. In *Advances in Neural Information Processing Systems*, pp. 585–593, 2009.

[7] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, Vol. 99, No. 12, pp. 7821–7826, 2002.

[8] Junzo Kamahara, Tomofumi Asakawa, Shinji Shimojo, and Hideo Miyahara. A community-based recommendation system to reveal unexpected interests. In *Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International*, pp. 433–438. IEEE, 2005.

[9] Da Kuang, Chris Ding, and Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 106–117. SIAM, 2012.

[10] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, Vol. 2, No. 1-2, pp. 83–97, 1955.

[11] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Vol. 401, pp. 788–791. Nature Publishing Group, 1999.

[12] Jialu Liu and Jiawei Han. Hinmf: A matrix factorization method for clustering in heterogeneous information networks. In *Proceedings of the international joint conference on artificial intelligence workshop*, 2003.

[13] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pp. 665–672, 2009.

[14] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494. AUAI Press, 2004.

[15] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 22, No. 8, pp. 888–905, 2000.

[16] Yizhou Sun, Yintao Yu, and Jiawei Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 797–806. ACM, 2009.

[17] Xiao Wang, Di Jin, Xiaochun Cao, Liang Yang, and Weixiong Zhang. Semantic community identification in large attribute networks. In *AAAI*, pp. 265–271, 2016.

[18] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 587–596. ACM, 2013.

[19] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining*, pp. 1151–1156. IEEE, 2013.

[20] Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. Latent community topic analysis: integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 3, No. 4, p. 63, 2012.

[21] Haizheng Zhang, Baojun Qiu, C Lee Giles, Henry C Foley, and John Yen. An lda-based community structure discovery approach for large-scale social networks. *ISI*, Vol. 200, , 2007.