DBSJ Japanese Journal Vol. 14-J, Article No. 1
March 2016

多重有向グラフのコア部抽出のための MDSR 法の提案と評価

Proposition and Evaluation of MDSR Method for Core Analysis of Multiple Directed Graphs

加藤 翔子♡ 斉藤 和巳♢
風間 一洋♠佐藤 哲司♣

Shoko KATO Kazumi SAITO
Kazuhiro KAZAMA Tetsuji SATOH

本稿では，多重有向グラフのコア部抽出手法として，**MDSR**（**Multiple-Directed-Spectral-Relaxation**）法を提案する．**MDSR** 法は，多重有向隣接行列の右固有ベクトルと左固有ベクトルを **2** 値に量子化してコア部を抽出する．さらに，隣接行列から抽出したコア部に含まれるリンクを削除した後に上記の処理を再帰的に適用することで，複数のコア部の抽出を実現する．左右の固有ベクトルを計算することで，リンクの始点ノード集合と終点ノード集合という，相補的な役割を果たす非対称な関係のノード群を組で抽出する．**Twitter** のリプライデータを用いた評価実験では，不特定多数にリプライを頻繁に送信するユーザ (例，**bot**)，リプライを頻繁に受信するユーザ (例，ユーザ名 **'null'**)，互いに頻繁にリプライを送り合う小規模グループのユーザという，顕著な行動を示す **3** タイプのユーザの存在を明らかにし，ノード群の重なりを持つ複数のコミュニティも個別に抽出できることを示す．また，単純無向グラフのコア部抽出手法である *k*-**core** 抽出法を多重有向グラフに向け拡張した **2** つの手法と比較し，**MDSR** 法の有効性を明らかにする．

In this paper, we propose the MDSR method for a problem of extracting core portions of a multiple directed network. The MDSR method extracts plural core portions by repeating the following two steps; quantizing elements of the left- and right-eigenvectors of an adjacency matrix of a network to binary ones as an indicator of extracting core portion, and removing links of the extracted one. By calculating the left- and right-eigenvectors, the MDSR method extracts pairs of asymmetric node sets which have complementary roles, i.e., initial and terminal nodes. In our experiments using a reply network on Twitter, we demonstrate that the MDSR method uncover the following three types of users; 1) users who send tweets frequently (e.g., bots), 2) users who receive tweets frequently (e.g., 'null'), 3) small groups who send/receive tweets frequently each other. We also show that some communities were overlapped ones. Furthermore, we show that such communities were hard to be automatically found by two methods, which were constructed by straightforwardly extending the conventional *k*-core method.

♡ 非会員　静岡県立大学大学院経営情報イノベーション研究科
445.ka10@gmail.com
♢ 正会員　静岡県立大学大学院経営情報イノベーション研究科
k-saito@u-shizuoka-ken.ac.jp
♠ 正会員　和歌山大学システム工学部
kazama@sys.wakayama-u.ac.jp
♣ 正会員　筑波大学図書館情報メディア系
satoh@ce.slis.tsukuba.ac.jp

1. はじめに

近年のスマートフォンの普及により，我々のコミュニケーション手段は大きく変貌し，特に Twitter や Facebook などに代表されるソーシャルメディアの成長は著しく，流通する情報量は増加し続けている [1, 2]．このようなソーシャルメディア内の友人関係で形成される社会ネットワークの基本構造や機能を理解するためにはコア部抽出が重要な役割を果たす [3]．しかし，そのリンクには方向性と重みが存在することが多いにもかかわらず，それを扱うことができなかった．この課題を解決するために，我々は，大規模な多重有向グラフのコア部を抽出できる MDSR (Multiple-Directed-Spectral-Relaxation) 法を提案し，Twitter のリプライ関係をデータセットとした評価実験により，その有効性を示した [4]．

既存のコミュニティ分析手法は大きく 2 つに大別できる．1 つは，グラフの結合が疎な部分を切断して，ノードを複数のクラスタに分類する手法 [5, 6, 7] であり，ノードが属するコミュニティは 1 つのみとなる．もう 1 つは，グラフ上で密結合したノード集合をコミュニティと見なして，クリークやクリークの条件を緩めたノード集合 (コア部) を抽出する手法 [8, 9, 10, 11] である．こちらは，ノードが複数のコミュニティに属することを許容するため，多様な人間関係から成る社会ネットワークの分析に適しており，MDSR 法もこれに含まれる．MDSR 法の構想は，隣接行列の固有ベクトルの観点からコア部を抽出する SR 法 [11] に基づくが，SR 法は単純無向隣接行列から計算される固有ベクトルを量子化するのに対し，MDSR 法では多重有向隣接行列から計算される，右側・左側の二つの固有ベクトルを量子化することで，互いに関係する二つの種類のノード群としてコア部を抽出できる点が異なる．

また，固有ベクトルの観点からネットワークを分析する既存手法として，HITS アルゴリズム [12] や PageRank アルゴリズム [13] が挙げられるが，これらはノードの重要度を求める手法であり，コア部抽出手法への拡張は容易でない．MDSR 法は，コア部抽出とその内部リンクの削除を再帰的に繰り返すことで複数コア部を自動的に抽出する．

本稿では，特に相補的な役割を果たす非対称な 2 種類のノード群を一つのコア部として抽出する MDSR 法の重要な特徴に注目して，[4] の内容を整理しなおすと共に，その特徴を明確化するために実データを使った評価実験を行う．

2. 多重有向グラフ分析手法

2.1 多重有向グラフ分析の必要性

近年のソーシャルメディアの成長に伴い，様々な大規模ネットワークが容易に入手できるようになった．このようなネットワークには重みや方向性があることが多く，それらを考慮した分析が必要性は以前から論じられてきた [14, 15]. しかし，これまで議論されてきた重要ノード決定アルゴリズムやコア部抽出手法の多くは，単純グラフや無向グラフを対象としてきたため，現実の現象を単純化できる利点はあるものの，必ずしも正しく分析できているとは言い難かった．ただし，従来法も多重有向グラフを扱えるように拡張できる．そこで，MDSR 法と比較するために，単純有向グラフの重要ノードを決定する HITS アルゴリズム [12] と単純無向グラフのコア部を抽出する k-core 抽出法 [10] を，以下のように多重有向グラフに向け拡張した．

2.2 HITS アルゴリズムの拡張

ネットワークの重要ノード探索手法として，Kleinberg らは HITS アルゴリズムを提案している．このアルゴリズムで多重有向グラフを扱うために，以下のように拡張する．

全ノード集合 $V = \{1, \cdots, N\}$，隣接行列 $\mathbf{A}$ から成る多重有向グラフ $G = (V, E)$ を考え，隣接行列の要素 (i, j) を $A(i, j)$ と定義する．$A(i, j)$ の値はノード $i \in V$ からノード $j \in V$ へのリンク本数であり，非負整数をとる．本稿では自己リンクは考慮しないため，$A(i, j) \neq A(j, i)$ かつ $A(i, i) = 0$ である．

このとき，$v(j) = \sum_{i \to j} u(i) = \sum_i A(i, j)u(i)$ と $u(i) = \sum_{j \to i} v(j) = \sum_j A(j, i)v(j)$ を満たすような $\mathbf{v} = \mathbf{Au}$ を Hub ベクトル，$\mathbf{u} = \mathbf{A}^T\mathbf{v}$ を Authority ベクトルとする．

以上の手法で多重有向グラフにおける重要ノードを探索できるが，先述したように，複数のコア部を抽出するように変更することは簡単ではない．

2.3 k-core 抽出法の拡張

代表的なコア部抽出法のひとつとして k-core 抽出法がある．この抽出法を強連結・弱連結の 2 つの視点から以下のように拡張する．

- コミュニティ定義

多重有向グラフ G の任意のノード $i \in V$ について，出次数を $d^+(i) = \sum_{j>0} A(i, j)$，入次数を $d^-(i) = \sum_{j>0} A(j, i)$ と定義する．本稿では，$d^+(i)$ と $d^-(i)$ が共に k 以上である強連結な部分グラフ $SC(k)$ を k-CSC コミュニティ，$d^+(i)$ と $d^-(i)$ の和が k 以上である弱連結な部分グラフ $WC(k)$ を k-CWC コミュニティと呼ぶ．

より詳細には，変数 k について，k-CSC コミュニティは以下のノード群 $V_{SC(k)} \subset V$ とリンク群 $E_{SC(k)} \subset E$ から構成される部分グラフ $SC(k) = (V_{SC(k)}, E_{SC(k)})$ である．

$$\begin{aligned} V_{SC(k)} &= \{i : d^+(i) \geq k, d^-(i) \geq k\}, \\ E_{SC(k)} &= \{e_{i,j} : i, j \subset V_{SC(k)}\}. \end{aligned} \quad (1)$$

同様に，k-CWC コミュニティは以下のノード群 $V_{WC(k)} \subset V$ とリンク群 $E_{WC(k)} \subset E$ から構成される部分グラフ $WC(k) = (V_{WC(k)}, E_{WC(k)})$ である．

$$\begin{aligned} V_{WC(k)} &= \{i : d^+(i) + d^-(i) \geq k\}, \\ E_{WC(k)} &= \{e_{i,j} : i, j \subset V_{WC(k)}\}. \end{aligned} \quad (2)$$

これを満たす最大部分グラフを $SC(k)$ と $WC(k)$ とみなし，その連結成分を k-CSC・k-CWC コミュニティとする．

- 抽出アルゴリズム

与えられたコア数 k について，k-CSC コミュニティをグラフ $G = (V, E)$ から抽出する．$(V_{SC(k)}, E_{SC(k)})$ を計算する基本手順を以下に示す．

A1. $V_{SC(k)} = V$, $E_{SC(k)} = E$ として初期化;
A2. $P = \{i : d^+(i) < k \ \ or \ \ d^-(i) < k\}$ である P を計算;
A3. $P = \emptyset$ なら $(V_{SC(k)}, E_{SC(k)})$ を出力し終了; でなければ $V_{SC(k)} = V_{SC(k)} - P$, $E_{SC(k)} = E_{SC(k)} - \{e_{i,j} : i \in P, j \in P\}$ として **A2**. へ戻る．

ここで P は k-CSC コミュニティから除外されるノード群である．同様に $(V_{WC(k)}, E_{WC(k)})$ も $P = \{i : d^+(i) + d^-(i) < k\}$ として **A2**. を再帰的に計算することで求まる．

なお，最適なコア数 k を前もって知ることは不可能であるが，k-CSC コミュニティ抽出の場合，最大コア数は $\max_i\{d^+(i)\}$ か $\max_i\{d^-(i)\}$ のいずれかであり，$SC(k+1) \subset SC(k)$ となることも考慮すれば，全ての k について再帰的に k-CSC コミュニティを計算する事が可能となる．

B1. $V_{SC(1)} = V$, $E_{SC(1)} = E$, $k = 2$ として初期化;
B2. $(V_{SC(k-1)}, E_{SC(k-1)})$ から $(V_{SC(k)}, E_{SC(k)})$ を計算;
B3. $V_{SC(k)} = \emptyset$ なら終了; でなければ $k = k + 1$ とし **B2**. へ戻る．

同様に，最大コア数を $\max_i\{d^+(i) + d^-(i)\}$ とすれば k-CWC コミュニティを抽出できる．この 2 つの抽出法を以下では k-CSC 法・k-CWC 法と呼ぶ．

3. MDSR 法

MDSR 法は，SR 法 [11] と同様に，Web のハイパーリンクで構成されるようなネットワークの比較的リンクが密集する部分に注目すべき構造が内在するとし，そのコア部を抽出する．実際には，隣接行列の固有ベクトルを求めてノードをランキングし，結合が密なコア部を抽出し，さらに抽出コア部のリンクを削除した後で再帰的に処理を繰り返すことで，複数のコア部を抽出する．MSDR 法における主たる拡張は次の 2 点である．

まず，SR 法は単純無向グラフを対象とするのに対し，MDSR 法では多重有向グラフを対象とするので，右側と左側の二つの固有ベクトルを考える．この右側固有ベクトルは 2.2 節にて言及した HITS アルゴリズムの Authority 度ベクトル，左側固有ベクトルは Hub 度ベクトルに対応する．例えば，Twitter のユーザ間のリプライ関係のような有向で回数の重みを持つネットワークの場合には，直感的には，右側固有ベクトル要素値の大きいノードは多くのリプライを受けるユーザ，左側は多くのリプライを出すユーザとなる．これらユーザ間に，Hub ノードから Authority ノードへの密なリンクが存在するならば，高いリンク多重度を持つ部分二部グラフ構造が存在することとなる．

次に，アルゴリズムの観点での特徴は，大規模グラフへの適用を可能にするために，固有ベクトル要素の量子化を右側と左側で独立に実行し，計算量を削減することである．このような独立実行の妥当性は，後述する実験で示すように右側・左側固有ベクトルが大きなギャップを有することに基づく．以下では，基本問題設定，緩和問題とその解法，基本問題の解法，および, 再帰的コア部抽出法の詳細に加えて，提案アルゴリズムの計算量は総リンク数の線形オー

ダであり，グラフがスパースなら大規模問題にも適用可能であることについて述べる．

3.1 基本問題設定

与えられたグラフ G の全ノード集合を $V = \{1, \cdots, N\}$ とし，その隣接行列を $\mathbf{A}$ とする．隣接行列の第 (i, j) 成分 $A(i, j)$ は，ノード i から j へ張られたリンクの本数を表す．本稿では，自己リンクなしの多重有向グラフを対象とするので，$A(i, i) = 0$ かつ $A(i, j)$ は非負整数となる．なお，自己リンクなしとする理由は，自分自身へ莫大な数のリンクを張るノードをコア部と呼ぶのは不自然なためである．

2 つのノード部分集合 $W \subseteq V$ と $X \subseteq V$ に対し，その間に張られた平均リンク数は以下となる．

$$G(W, X) = \frac{1}{\sqrt{|W||X|}} \sum_{i \in W} \sum_{j \in X} A(i, j). \tag{3}$$

$|W|$ は集合 W の要素数を表す．リンクが密集する部分に注目すべき構造があると既に述べたが，MDSR 法ではこの構造が部分集合ペア W と X として形成されると想定する．そこで，式 (3) を最大にするノード部分集合ペア W と X の探索問題を考える．ただし，単純な数え挙げによる網羅的な探索では，N の大きなグラフでは組合せ爆発が容易に起こる．よって，MDSR 法では，以下で述べるように緩和問題が最適に解けることに着目したアプローチを採用する．

3.2 緩和問題とその解法

ノード部分集合 W に対して，N 次元ベクトル $\mathbf{q}$ を，$i \in W$ ならば $q(i) = 1$，さもなければ $q(i) = 0$ で定義する．同様に，ノード部分集合 X に対して，N 次元ベクトル $\mathbf{r}$ を定義する．このとき，式 (3) は以下のように書き換えられる．

$$G(\mathbf{q}, \mathbf{r}) = \frac{\mathbf{r}^T \mathbf{A} \mathbf{q}}{\sqrt{\mathbf{q}^T \mathbf{q} \mathbf{r}^T \mathbf{r}}}. \tag{4}$$

ただし，$\mathbf{r}^T$ はベクトル $\mathbf{r}$ の転置を表す．ここで，ベクトル $\mathbf{q}$ の各要素に対して連続値まで許容すれば，$G(\mathbf{q}, \mathbf{r})$ の最大値は，行列 $\mathbf{A}$ に対して，右側と左側の固有ベクトル $\mathbf{q}^*$ と $\mathbf{r}^*$ で与えられる．この $\mathbf{q}^*$ と $\mathbf{r}^*$ が 2.2 節で言及した Authority 度ベクトル $\mathbf{u} = \mathbf{A}^T \mathbf{v}$ と Hub 度ベクトル $\mathbf{v} = \mathbf{A}\mathbf{u}$ に対応する．

固有ベクトル $\mathbf{q}^*$ と $\mathbf{r}^*$ を求めるために，以下のパワー法を土台としたアルゴリズムを適用する．

E1. $t = 1$, $\mathbf{q}^{(0)} = (1, \cdots, 1)^T$ と初期化する;
E2. $\widetilde{\mathbf{q}} = \mathbf{A}^T \mathbf{A} \mathbf{q}^{(t-1)}$, $\mathbf{q}^{(t)} = \widetilde{\mathbf{q}} / \max_i \widetilde{q}(i)$ を求める;
E3. $\max_i |q^{(t)}(i) - q^{(t-1)}(i)| < \epsilon$ なら反復を終了;
E4. $t = t + 1$ として **E2** に戻る．

ここで，ϵ は終了条件を制御する正の実数であり，反復終了後に $\mathbf{q}^* = \mathbf{q}^{(t)}$，および，$\mathbf{r}^* = \mathbf{A}\mathbf{q}^*$ として結果が求まる．明らかに，$\mathbf{A}$ と $\mathbf{q}^{(0)}$ の全要素が非負のため，任意の反復で $\widetilde{\mathbf{q}}$ の各要素は非負となる．さらに，**E2** でスケーリングを施すことより，$0 \le q^{(t)}(i) \le 1$ が保証される．よって，上記アルゴリズムで右側と左側の固有ベクトルのペアが求まり，基本問題設定の妥当な緩和問題を最適に解くことができる．

上記アルゴリズムの 1 反復の主要計算量は，グラフの総リンク数を L とすれば，$\widetilde{\mathbf{q}}$ は高々 L 回の掛け算と足し算で求まる．一方，$\widetilde{\mathbf{q}}$ のスケーリングは，ノード数 N に対し，計算量 $O(N)$ の乗算で実現できる．

3.3 量子化問題とその解法

固有ベクトル $\mathbf{q}^*$ と $\mathbf{r}^*$ を個別に考え，各要素をバイナリ化することで基本問題の解を求める．以下では，$\mathbf{q}^*$ の量子化法について述べる．なお，$\mathbf{r}^*$ も同じ手順で量子化できる．まず，$\mathbf{q}^*$ の要素の大小に基づき各ノードをランキングすれば，リスト $S = [s(1), \cdots, s(N)]$ が定まる．ここで，$s(i)$ はランク i に対して元のノード番号を与える関数で，$q^*(s(i)) \ge q^*(s(i+1))$ の関係を満たす．なお，tie-break は任意に行なうとする．いま，ベクトルの全要素を単一の値で量子化するならば，二乗誤差 $E(0)$ は平均値で最小化され，$E(0)$ は式 (5) で求まる．

$$E(0) = \sum_{i=1}^{N} (q(i) - \frac{1}{N} \sum_{j=1}^{N} q(j))^2 = \sum_{i=1}^{N} q(i)^2 - \frac{1}{N} (\sum_{i=1}^{N} q(i))^2. \tag{5}$$

次に，リスト $S = [s(1), \cdots, s(N)]$ の上位 m 個のノード集合 $W(m)$ と下位の $N - m$ 個のノード集合で量子化するならば，前後それぞれの区間での平均値により二乗誤差 $E(m)$ が最小化され，式 (7) のように計算できる．

$$\begin{aligned} E(m) &= \sum_{i=1}^{m} (q(s(i)) - \frac{\sum_{j=1}^{m} q(s(j)))^2}{m} \\ &\quad + \sum_{i=m+1}^{N} (q(s(i)) - \frac{\sum_{j=m+1}^{N} q(s(j)))^2}{N-m} \\ &= \sum_{i=1}^{N} q(i)^2 - \frac{1}{m} (\sum_{i=1}^{m} q(s(i)))^2 - \frac{1}{N-m} (\sum_{i=m+1}^{N} q(s(i)))^2 \end{aligned} \tag{6}$$

MDSR 法では，式 (7) の $E(m)$ を最大にする m^* を探索し，ノード集合 $W(m^*)$ を求める．効率良く m^* を探索するため，ソートした固有ベクトルの要素値 $(q(s(1)), \cdots, q(s(N)))$ を次式で $(y(0), y(1), \ldots, y(N))$ へと変形する．

$$\begin{aligned} y(0) &= 0, \\ y(i) &= \sum_{j=1}^{i} q(s(j)) = y(i-1) + q(s(i)) \quad (i = 1, \cdots, N). \end{aligned} \tag{7}$$

このとき，二乗誤差 $E(m)$ は次式で求まる．

$$E(m) = \sum_{i=1}^{N} q(i)^2 - \frac{1}{m} y(m)^2 - \frac{1}{N-m} (y(N) - y(m))^2. \tag{8}$$

上記手順を以下に整理する．

F1. $\mathbf{q}^*$ の要素をソートしランク関数 $s(i)$ を求める;
F2. $(q(s(1)), \cdots, q(s(N)))$ から $(y(1), \ldots, y(N))$ を式 (7) で求める;
F3. $E(1), \cdots, E(N-1)$ を式 (8) で求める;
F4. $m^* = \arg\max_m E(m)$ を求めて $W(m^*)$ を出力する;

上記アルゴリズムの主要計算量は以下となる．**F1** のソートは $O(N \log N)$ の計算量で実行できる．**F2** の変形は式 (7) における N 回の加算で実行でき，**F3** で $E(1), \cdots, E(N-1)$ を求めるには，$q(i)$ の二乗和を予め計算しておけば，式 (8) は $O(N)$ の計算量で求めることができる．

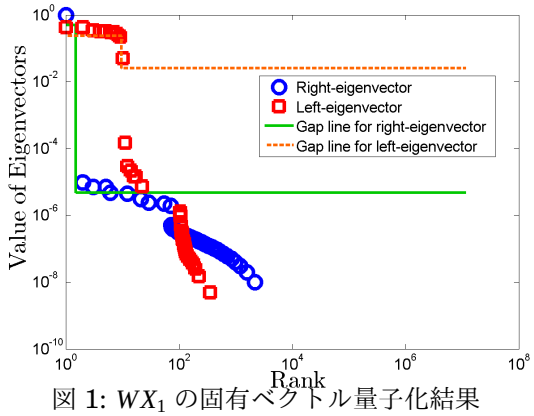

図 1: WX_1 の固有ベクトル量子化結果

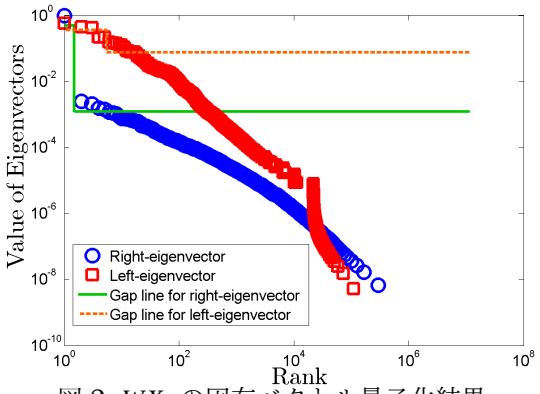

図 2: WX_2 の固有ベクトル量子化結果

3.4 コミュニティ抽出アルゴリズム

上述した処理手順を T 回繰り返すことで，結合が密な T 個のコア部を抽出する．全体のアルゴリズムを以下に示す．

G1. $t = 1$ から T まで以下のステップを実行する;
G2. **E1** から **E4** を反復させ $\mathbf{q}_m^*$ を求める;
G3. $\mathbf{q}^*$ に対し **F1** から **F4** で $W_k(m^*)$ を求める;
G4. $\mathbf{r}^*$ に対し **F1** から **F4** で $X_k(n^*)$ を求める;
G5. $i \in W_k(m^*), j \in X_k(n^*)$ なら $A(i, j) = 0$ とする．

ここで，抽出コア数 T はユーザが任意に設定するパラメータであり，最終的な結果は，コア部として T 個の集合ペア $(W_1^*, X_1^*), \cdots, (W_T^*, X_T^*)$ が求まる．

4. 評価実験

Twitter のリプライをデータセットとし，MDSR 法による抽出結果を k-CSC·CWC 法と比較する．さらに，MDSR 法が複数のコミュニティに属するノードも抽出できることを示す．なお，以下では MDSR 法により t 回目で抽出されたコア部 (W_t^*, X_t^*) を WX_t と表す．この W_t^* はリプライ受信数，X_t^* はリプライ送信数が優位なノード群と捉えることができる．また，実験では $T = 100$ としてコア部を抽出したが，本稿では $1 \le t \le 10$ の結果を紹介する．

4.1 データセット

本稿では，2012 年 3 月 14 日から 2013 年 3 月 14 日の期間に，日本国内で投稿された Twitter のツイート [16, 17] のうち，文頭が '@screen_name' 形式で始まるリプライツイートをデータセットとした．ユーザをノードとし，各ユーザから '@screen_name' 形式で指定されたユーザへリンクを張ることで，多重有向ネットワークを構築した．ネットワークのノード数は 11,500,369，多重リンク数は 1,649,048,139 であり，リンクの平均多重度はおよそ 12.8 であった．

表 1: MDSR 法により抽出した 10 コアの概要

WX_t	$\lvert W_t^*\rvert$	$\lvert X_t^*\rvert$	accounts in W	accounts in X
WX_1	1	9	8∗[1]	8∗[2...6], hir∗[1...4]
WX_2	1	5	null	gat∗, yuu∗, . . .
WX_3	1	1	113∗	e21∗
WX_4	7	2	toa∗[1...4], mik∗, . . .	toa∗[1,2]
WX_5	1	1	Ten∗	Key∗
WX_6	4	1	kyo∗[1...4]	Ya_∗
WX_7	8	1	ziz∗, dr8∗, . . .	dq_∗
WX_8	1	1	pos∗	S_c∗
WX_9	3	1	Sox∗, car∗, Ara∗	mom∗
WX_{10}	1	19	null	chi∗, miy∗, . . .

4.2 固有ベクトルの量子化結果

量子化問題の解法の妥当性を確認するために，各コア部の多重隣接行列の右固有ベクトル·左固有ベクトルを量子化した結果を示す．図 1，2 は，WX_1, WX_2 の抽出時の固有ベクトルの値と量子化点のプロットである．青色の丸いマーカは右固有ベクトル，赤色の四角いマーカは左固有ベクトルの要素の値を表し，緑色の実線は右固有ベクトル，橙色の破線は左固有ベクトルを 2 値に量子化した点を表す．直感的には，実線や破線の垂直部より左に位置する固有ベクトル要素値を持つノードをコア部として抽出している．

図 1 から，WX_1 の抽出時点では，固有ベクトル要素値の高いノードと低いノードで明確に 2 分されていることが確認できる．$t \ge 3$ の抽出結果でも，ほとんどがこのように 2 分されており，量子化問題の解法の妥当性が確認できる．

一方，図 2 から，WX_2 の抽出時点では，他の結果に比べて左固有ベクトル要素値の高低差が小さい．同様の結果は $t \ge 3$ での抽出結果でもまれに見られる．このような場合，W_t^* として抽出されたノードは，他のコア部としても重複して抽出されることを確認している．例えば，$t = 2$ で W_2^* として抽出されたノードは，$t = 10$ でも W_{10}^* として抽出されている．詳細は 4.3 節に示す．

4.3 抽出結果：MDSR 法

MDSR 法によりコア部抽出されたノードの特徴を把握し，MDSR 法によるコア部抽出の妥当性を確かめるために，ノード数とユーザ名を調査する．表 1 は，$1 \le t \le 10$ での抽出コア部 WX_t における，ノード数と構成ユーザの調査結果である．なお，$|W_t^*|$ は W_t^* として抽出されたノードの個数であり，$|X_t^*|$ についても同様である．また，ユーザ名については，先頭の数文字とワイルドカード文字 $*$ を使ったパターンで表し，先頭文字列が同じユーザが複数存在する場合は abc∗[1] のように番号を付加して区別する．

表 1 より，WX_1，WX_4，WX_6 では，同じ文字列から始まるユーザ名を共通して持つノードがコア部として抽出されている．これらのノードはプログラムにより機械的に作成されたユーザであると考察でき，MDSR 法はこのようなユーザ群をまとめて一つのコミュニティとして抽出することがわかる．また，WX_2，WX_{10} では，X_2^* と X_{10}^* として 'null' というユーザが重複して抽出されている．Twitter では，'null' はユーザ名としてではなく，発言の文頭に '@null' を指定す

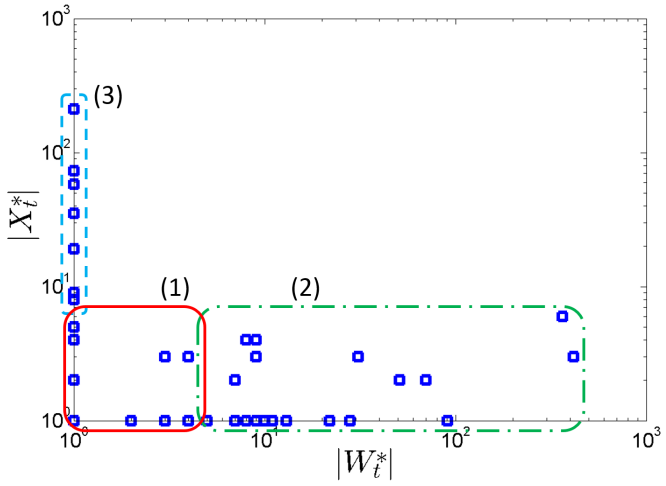

図 3: $|W_t^*|$ と $|X_t^*|$ の分布

ることでフォロワーのタイムラインに自分の発言を表示させないために使われている．このように‘null’が X_t^* として抽出されているコミュニティは $t \geq 11$ でも確認できた．

更に，100 個の抽出コア部の $|W_t^*|$ と $|X_t^*|$ の分布を調査し，図 3 にプロットした．なお，$|W_t^*|$ の値が大きく $|X_t^*|$ の値が小さいコミュニティや，$|W_t^*|$ の値が小さく $|X_t^*|$ の値が大きいコミュニティは存在するが，$|W_t^*|$ と $|X_t^*|$ の値が同程度のコミュニティは，(1,1) や (2,2) など規模が小さいものに限られ，共に大きな値をとる WX_t は $1 \geq t \geq 100$ では確認できなかった．一般に，フォロー数・被フォロー数や受信・送信リプライメッセージ数に存在するユーザごとの大きな偏りが，抽出される $|W_t^*|$ と $|X_t^*|$ の大きさの偏りの原因だと推測する．したがって，$|W_t^*|$ と $|X_t^*|$ の比によって，(1) 赤色の実線で囲まれた $|W_t^*| \simeq |X_t^*|$ であるコア部，(2) 緑色の鎖線で囲まれた $|W_t^*| > |X_t^*|$ であるコア部，(3) 水色の破線で囲まれた $|W_t^*| < |X_t^*|$ であるコア部の 3 種類に分類できる．

この 3 分類について，コア部を構成するユーザを調査した．(1) に分類されるコア部では，少人数の親密なグループ内のメッセージ交換が見られた．(2) に分類されるコア部では，$|X_t^*|$ を構成するユーザは，フォロワーの通常ツイートに反応して自動的に reply を送る bot アカウントが多く見られた．(3) に分類される $|W_t^*| = 1$ のコア部では，$|W_t^*|$ を構成するユーザは‘null’や YouTube[1] の公式アカウントなどが見られた．YouTube の公式アカウントも‘null’と同様に，特定のユーザ名として使われることは少なく，YouTube の動画再生ページの共有機能を使って Twitter 上で動画を紹介する際に，デフォルトのツイート本文に‘@YouTube’が含まれるためと推測される．このようなコア分類は，抽出コア部が W_t^* と X_t^* の 2 つのノード群で構成される MDSR 法特有のものである．

4.4 抽出結果：k-CSC 法・k-CWC 法

MDSR 法の有効性を確かめるために，k-CSC 法・k-CWC 法で抽出したコア部のノード数とユーザ名を調査し，MDSR 法と比較する．表 2，3 は，k の降順でソートした上位 10 コアのノード数と構成ユーザの調査結果であり，ソート後の順位 r に沿って SC_r，WC_r として表記する．なお，$|V_{SC(k)}|$ は $SC(k)$ として抽出されたノードの個数であり，$|V_{WC(k)}|$ についても同様である．また，構成ユーザが MDSR 法でも抽出されている場合は，WX_t として対応するコア部を示す．

[1]http://www.youtube.com/

表 2: k-CSC 法により抽出した 10 コアの概要

SC_r	k	$\lvert V_{SC(k)}\rvert$	accounts	WX_t
SC_1	35184	2	toa∗[1,2]	WX_4
SC_2	22388	2	sen∗, get∗	-
SC_3	22010	2	twi∗, non∗	-
SC_4	18640	2	ant∗, hha∗	-
SC_5	10745	2	ats∗, Not∗	-
SC_6	10352	2	rks∗, gom∗	-
SC_7	9948	2	sou∗, 287∗	-
SC_8	9599	3	AOI∗, bot∗[1,2]	-
SC_9	9423	2	U_S∗, h_t∗	-
SC_{10}	9124	2	God∗, eru∗	-

表 3: k-CWC 法により抽出した 10 コアの概要

WC_r	k	$\lvert V_{WC(k)}\rvert$	accounts	WX_t
WC_1	110052	2	113∗, e21∗	WX_3
WC_2	91941	2	Ten∗, Key∗	WX_5
WC_3	76086	3	toa∗[1...3]	WX_4
WC_4	73394	2	pos∗, S_c∗	WX_8
WC_5	66623	2	null, gat∗	WX_2
WC_6	65458	4	toa∗[1...3], mik∗	WX_4
WC_7	58978	2	8∗[1], hir∗[1]	WX_1
WC_8	58384	5	toa∗[1...4], mik∗	WX_4
WC_9	57740	3	8∗[1], hir∗[1,2]	WX_1
WC_{10}	52342	2	Sox∗, mom∗	WX_9

表 2 より，k-CSC 法は頻繁にリプライを送り合う 2 者を抽出する傾向があることがわかる．また，表 3 より，k-CWC 法により抽出された上位 10 コアは，MDSR 法で抽出されたコア部の一部であることがわかる．より詳細には，WC_3 は WC_6 の，WC_6 は WC_8 の部分グラフであり，いずれも WX_4 の部分グラフである．同様に，WC_7 は WC_9 の部分グラフであり，いずれも WX_1 の部分グラフである．

この結果より，k-CWC 法は，k の値を下げることで妥当な大きさのコミュニティを抽出し得ることが示唆される．しかし，k の最適値の決定に関しては，現時点では適切な自動化手法はなく，人間が結果を見て決定していることから，ソーシャルメディアのような最大次数が高いネットワークほど調査する k の値の範囲が広くなり，作業は困難になる．これに対し，MDSR 法は妥当な大きさのコミュニティを自動的に抽出できる．さらに，k-CSC 法・k-CWC 法は，ノード間のリンク構造が一見しただけではわからないが，MDSR 法は抽出コア部を W_t^* と X_t^* の二つのノード群に分けて俯瞰できるため，容易にリンク構造が理解できる．したがって，MDSR 法は大規模多重有向グラフのコア部抽出手法として有効であると言える．

5. おわりに

大規模多重有向グラフのコア部抽出手法として MDSR 法を提案した．本手法は，多重コミュニティを個別に抽出できることに加えて，相補的な役割を果たす非対称な 2 種類のノード群を一つのコア部として抽出することから，新たなネットワーク分析が可能になることを明らかにした．

Twitter のリプライ関係をデータセットとして，多重性・有向性を考慮した結果，以下の知見を得た．まず，多重性を考慮したことで，1) リプライを頻繁に送信するユーザ，2) リプライを頻繁に受信するユーザ，3) 互いに頻繁にリプライを送り合う少数グループという，顕著な行動を示す 3 タイプのユーザの存在を明らかにした．さらに，k-core 法を多重有向グラフへ向け自然拡張した 2 手法と比較し，MDSR 法が自動で妥当な規模のコミュニティを抽出できることに加えて，相補的な役割を果たす非対称なノード群の間の関係を明らかにできることを示した．

本稿では Twitter 上の社会ネットワークをコア部抽出の対象としたが，MDSR 法は，単語-文書 2 部グラフや blog のトラックバックネットワークなど，様々なデータセットへ応用可能である．今後は，単純有向グラフなど他のネットワークデータを用いて定量的に評価し，様々なコア部抽出手法と結果を比較することで，MDSR 法の有効性をより明確化させる予定である．

[謝辞]

本研究は JSPS 科研費 25280110 の助成を受けた．

[文献]

[1] Huberman, B. A., Romero, D. M., Wu, F.: “Social networks that matter: Twitter under the microscope”, First Monday 14, (1.5). (2009)

[2] Kwak, H., Lee, C., Park, H., Moon, S.: “What is Twitter, a social network or a news media?”, In: Proceedings of WWW’10. pp. 591–600 (2010)

[3] Newman, M. E. J., Park, J.: “Why social networks are different from other types of networks”, *Physical Review E*, 68(3), 036122 (2003)

[4] Kato, S., Saito. K., Kazama, K., Satoh, T.: “MDSR: An Eigenvector Approach to Core Analysis of Multiple Directed Graphs”, In *PRICAI 2014: Trends in Artificial Intelligence*. Springer International Publishing, pp. 447-458 (2014)

[5] Shi, R., Malik, J.: “Normalized cuts and image segmentation”, IEEE Trans. PAMI, 22(8), pp.888–905 (2000)

[6] Flake, G. W., Lawrence, S., Giles, C. L.: “Efficient identification of Web communities”, In: Proceedings of SIGKDD’00. pp.150–160, (2000)

[7] Girvan, M., Newman, M. E. J.: “Community structure in social and biological networks”, In: Proceedings of the National Academy of Sciences of the United States of America, 99, pp.7821–7826 (2002)

[8] Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: “Uncovering the overlapping community structure of complex networks in nature and society”, Nature, Vol.435, pp.814–818 (2005)

[9] Saito, K., Yamada, T., Kazama, K.: “Extracting Communities from Complex Networks by the k-Dense Method”, IEICE Transactions, Vol.E91-A, No.11, pp.3304–3311 (2008)

[10] Seidman, S. B.: “Network structure and minimum degree”, Social Networks, Vol.5, No.3, pp.269–287 (1983)

[11] Saito, K., Ueda, N.: “Filtering Search Engine Spam based on Anomaly Detection Approach”, In: Proceedings of the KDD2005 Workshop on Data Mining Methods for Anomaly Detection, pp.62–67 (2005)

[12] Kleinberg, J.: “Authoritative sources in a hyperlinked environment”, Journal of the ACM, Vol.46, No.5, 604–632 (1999)

[13] Brin, S., Page, L.: “The anatomy of a large scale hypertextual Web search engine”, In: Proceedings of WWW’98, pp.107–117 (1998)

[14] Leicht, E. A., Newman, M. E.: “Community structure in directed networks”, *Physical review letters*, 100(11), 118703 (2008)

[15] Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: “The architecture of complex weighted networks”, *Proceedings of the National Academy of Sciences of the United States of America*, 101(11), pp.3747–3752 (2004).

[16] Yamamoto, S., Satoh, T.: “Two Phase Extraction Method for Extracting Real Life Tweets using LDA”, In: Proceedings of APWeb’13, pp.340–347 (2013)

[17] Yamaguchi,Y., Yamamoto, S. and Satoh, T.: “Behavior analysis methods for Twitter users based on transitions in posting activities”, Journal of Web Information Systems, Vol.10, No. 4, pp. 363–377, Emerald (2014)

加藤 翔子 Shoko KATO
2015 静岡県立大学大学院経営情報イノベーション研究科修士課程卒業，複雑ネットワークの研究に従事．同年 (株) トライバルメディアハウス入社．

斉藤 和巳 Kazumi SAITO
静岡県立大学大学院経営情報イノベーション研究科教授．1985 慶応義塾大学理工学部数理学科数学専攻卒業，同年日本電信電話 (株) 入社．1998 東京大学 博士 (工学)．複雑ネットワーク，機械学習などの研究に従事．情報処理学会，電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会各会員．

風間 一洋 Kazuhiro KAZAMA
和歌山大学システム工学部教授．1988 京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話 (株) 入社．2005 京都大学大学院情報学研究科システム科学専攻博士課程修了．博士 (情報学)．Web 情報検索，Web マイニングなどの研究に従事．情報処理学会，人工知能学会，日本ソフトウェア科学会，ACM 各会員．

佐藤 哲司 Tetsuji SATOH
筑波大学図書館情報メディア系教授．1980 山梨大学工学部電子工学科卒業．同年日本電信電話後者武蔵野電気通信研究所に入所．1994 大阪大学 博士 (工学)．論理回路の大規模一括集積技術，データベースマシン，マルチメディアデータベースの研究・開発に従事．電子情報通信学会，情報処理学会，ACM 各会員．