Twitter 上の arXiv からの学術情報流通に関する分析

嶋田　恭助† 　風間　一洋† 　吉田　光男†† 　佐藤　翔†††

† 和歌山大学システム工学部　〒 640–8510 和歌山県和歌山市栄谷 930
†† 豊橋技術科学大学情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
††† 同志社大学免許資格課程センター　〒 602–8580 京都市上京区今出川通烏丸東入
E-mail: †{s216327,kazama}@wakayama-u.ac.jp, ††yoshida@cs.tut.ac.jp, †††min2fly@slis.doshisha.ac.jp

あらまし　近年，学術論文に存在した有料の壁や情報公開の遅延などの問題解決手段として arXiv などのプレプリントサービスが注目されており，その普及に Twitter が重要な役割を果たしていると報告されているが，まだ定量的な分析はほとんどない．本論文では，プレプリントサービスを発信源とする学術情報の流通において，ユーザには情報の拡散と収集の 2 種類の役割があるとみなして，Twitter 上のユーザの行動から，その相補的な関係を分析する．具体的には，arXiv の URL の紹介ツイートとそれに対するいいね，リツイートの 3 種類の行動からユーザをノードとするグラフ構造を作成し，HITS アルゴリズムを用いて学術情報の拡散と収集の役割を果たしている主要ノード群を，それぞれ Authority と Hub として抽出する．さらにそれらのノード群をユーザや論文に関する属性情報を含めて分析することで，学術情報流通において人間や bot がどのような役割を果たしているかを明らかにする．

キーワード　Twitter，arXiv，学術情報流通，HITS アルゴリズム，Louvain 法

1 はじめに

Web の発展と共に，従来は紙ベースだった論文の電子化が進み，ACM Digital Library などの電子アーカイブや ELSEVIER などの出版社からインターネットを通じて容易に入手可能となった．さらに，Clarivate Analytics の Web of Science（WoS）や Google Scholar などのサービスを使った横断的な検索も可能になった．ただし，レベルの高い論文は少数の商業出版社による寡占状態にある [1] ために，必ずしも論文の全文を自由に閲覧できるわけではなく，さらにその論文購読料の上昇により購読契約を打ち切らざるをえない大学も出てきている状況 [1] を踏まえて，査読前の論文や最新の研究成果の公開を目的とした arXiv などのプレプリントサーバが注目され，研究情報の交換や学会の運営に積極的に活用されはじめている．

同時に研究者同士のコミュニケーション方法も，学会開催時などの限られた機会の対面コミュニケーションだけでなく，Twitter を使って随時議論するように変化し，学術情報もこれらのサービスを介して流通するようになった [2]．すなわち，重要あるいは新しいプレプリントの情報は，まず該当分野を常に調査している専門的なユーザや公式 bot などにより拡散され，その情報を受け取ったユーザが重要だと思った場合にはいいねして後から参照できるようにしたり，リツイートして再びフォロワーに拡散することで，広範囲に広まっていくと考えられる．

ただし，Twitter 上の学術情報流通の解析においては，いくつかの困難な課題が存在する．公式リツイートによる情報拡散では，正しい情報流通経路を知ることはできず，同一のユーザであっても，自らのツイートやリツイートによる情報拡散と，それらを閲覧することによる情報取得という 2 種類の役割を担っている可能性があり，学術情報流通を明らかにする場合，その二面性を分析する必要がある．arXiv では新着情報を拡散する bot が提供されており，必ずしもフォロワー数が多くはないものの，情報流通の起点として重要な役割を果たしていると考えられることから，従来の多くのソーシャルメディア分析 [3] [4] のように，単純に bot を除去する方向で分析するのではなく，人間と bot を包括して分析する仕組みが必要とされる．

本論文では，プレプリントサービスを発信源とする学術情報の流通において，既存研究のように情報流通経路を何らかの手法で推定するのではなく，Twitter 上のツイート，リツイート，いいねの 3 種類の行動から，ツイートではなくユーザをノードとするグラフ構造を作成し，ユーザには情報の拡散と集取の 2 種類の役割があるとみなして，その相補的な関係を分析する．具体的には，arXiv の URL の紹介ツイートとそれに対するいいね，リツイートの 3 種類の行動からグラフを作成し，HITS アルゴリズムを用いて学術情報の拡散と収集の役割を果たしている主要ノード群を，それぞれ Authority と Hub として抽出する．さらにそれらのノード群をユーザや論文に関する属性情報を含めて分析することで，学術情報流通において人間や bot がどのような役割を果たしているかを明らかにする．

2 関連研究

2.1 arXiv プレプリントと査読付き論文の関係の分析

arXiv ユーザは arXiv にプレプリントを投稿後に，査読を経て雑誌や国際会議に論文を発表すると考えられるが，そのような利用形態に関する学術情報 DB を用いた分析が存在する．

Larivière らは，arXiv と WoS の 2 つのデータソースを分析して，1991 年から 2011 年の間の arXiv プレプリントの約 64% が WoS に収録されており，その 93 %の分野は数学や物理学，

1：https://www.nature.com/articles/d41586-019-00758-x

地球科学，宇宙科学であることを明らかにした [5]．特に天文学，天体物理学，核物理学，素粒子物理学においては，WoS に収録された論文の大部分が arXiv にもプレプリントとして投稿されていた．数学，物理学も arXiv プレプリントを投稿している割合が高かったが，細かい分野で部分的に低いこともあった．

ただし，WoS は計算機科学分野の論文の収録割合が少ないことや，この分析ではその後急速に発展した AI 分野の利用状況は対象になっていないことに注意する必要がある．

2.2 arXiv プレプリントの影響の分析

arXiv プレプリント投稿後にソーシャルメディア上でその URL を含む発言が拡散すれば，より多くのユーザに閲覧されるようになると考えられる．そこで，論文の重要性を示す引用数やダウンロード数とソーシャルメディアにおける発言との関連性を分析することで，プレプリントが与える影響度や，逆にそれから論文の重要性を推定する研究が存在する．

Shuai らは，2010 年 10 月から 2011 年 5 月の間の 4,606 件のプレプリントに関して，Twitter 上の言及数と arXiv のダウンロード数，論文の引用数の関係を分析した [6]．その結果，言及数は，プレプリント投稿から数か月後の arXiv のダウンロード数と早期の引用数と中程度の相関があることを明らかにした．arXiv.org からダウンロードされ，Twitter でも言及されたほとんどのプレプリントの分野は，天体物理学や高エネルギー物理学，数学であり，70 %近い論文は投稿から 5 日以内に言及数のピークに達していることがわかった．ただし，Twitter データをランダムに 10%サンプリングする Twitter 社の Gardenhose Streaming API [7] で収集したために，Twitter 空間全体の反応を分析したわけではない．

また，Haustein らは，物理学や数学，CS 分野では論文よりも arXiv プレプリントに対してツイートされることが多い [6] [8] ことから，arXiv や WoS，Altmetrics データを用いて学術誌の論文と arXiv プレプリントに関する Twitter 上のアクティビティの調査を行った結果，例えば高エネルギー物理学の Twitter 言及数は bot の影響があることを示している [9]．

すなわち，ソーシャルメディアデータ上の学術情報流通では，bot の影響を含む固有の特性があることに注意が必要である．

3 学術情報流通の分析手法

3.1 Twitter 上の学術情報流通

本論文では，プレプリントサービス arXiv から発信される学術情報が，ソーシャルメディア上でどのように拡散・収集されているかを分析することで，学術情報が実社会に与えている影響や，学術情報の普及や利用に対するプレプリントサービスの貢献を明らかにするとともに，学術情報の多面的な側面を考慮した評価尺度を確立することを目的とする．

そのような学術情報流通においては，ソーシャルメディアサービスの中で Twitter が特に重要な役割を果たしていることが指摘されている．従来は，ソーシャルメディア上の情報流通は，ユーザあるいは発言が繋がったグラフ上を発信された情報が伝播した経路を検出または推定することで分析されていた．

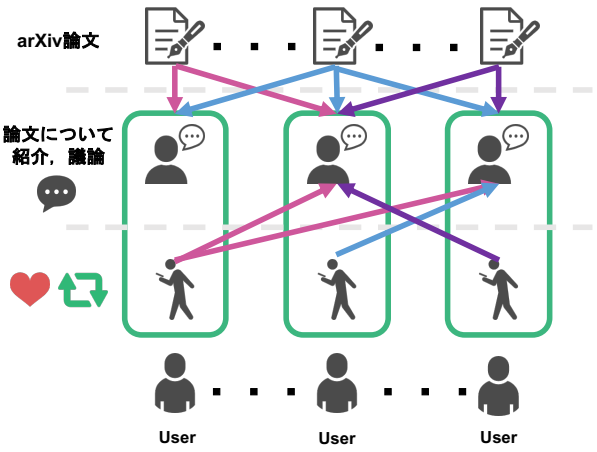


図 1: 学術情報流通の 3 層モデル

ただし，Twitter における学術情報流通を分析しようとした場合には，いくつかの課題がある．まず，Twitter システムの外部から観測できるツイート，リツイート，いいねなどのユーザの行動から情報拡散を分析する場合には，あるツイートを誰がリツイート・いいねしたかはわかっても，タイムラインのどのツイートまたはリツイートに触発されたかまでは判別できず，それを何らかの方法で推定したとしても，情報伝播経路の推定を繰り返すことで，現実との乖離が大きくなる可能性がある．次に，Twitter 上のユーザには，ツイートやリツイートによる情報拡散と，それらの閲覧による情報取得という 2 種類の役割があると考えられることから，情報流通におけるユーザの性質を把握するためには，両方の役割を考慮して分析する必要がある．最後に，arXiv では新着情報を拡散する bot が公式に提供され，必ずしもフォロワー数は多くないが情報流通の起点として重要な役割を果たしていると考えられることから，従来のソーシャルメディア分析のように，単に bot を判定・除外せずに，人間と bot を包括して分析する仕組みが必要とされる．

3.2 学術情報流通の 3 層モデル

本論文では，Twitter 上の arXiv からの学術情報流通を，ソーシャルメディアにおけるユーザの役目という観点から図 1 に示すような 3 層としてモデル化する．これを学術情報流通の 3 層モデルと呼ぶ．1 層目は arXiv 論文であり，学術情報はその URL として Twitter 上を流通する．なお，ここで arXiv プレプリントではなく，arXiv 論文と呼ぶ理由は，他のデータベースなどを使ってプレプリントだけに絞り込んでいいないからである．2 層目は情報拡散者（information spreader），3 層目は情報収集者（information collector）を表す．ここで上下の関係にある情報拡散者と情報収集者が四角形で囲まれているのは，同一のユーザが情報拡散と情報取得という異なる役目を同時に持つことができることを表す．まず，情報拡散者は，arXiv 論文の URL をツイート，またはそれをリツイートすることで，学術情報を拡散させる．さらに，情報収集者は，その情報を取得し，さらに価値があると判断した場合に限り，そのツイート

をリツイートしたり，いいねしたりする．ただし，あるユーザがコメント付きでリツイートし，それに対してリツイートまたはいいねされた場合に限り，そのユーザは情報収集者であると共に情報拡散者となる．もちろん，ユーザによっては，どちらか片方しかおこなわないこともありえる．

本モデルでは，Twitter 上のフォロー・非フォロー関係にあるユーザ同士の同類性が低いという知見から，情報が拡散または収集されるかは，それを拡散したユーザの性質ではなく，情報の内容で判断されると仮定する．さらに，一般的なニュースと異なり，arXiv プレプリントのような学術情報は刻々と内容が変わりながら伝播することはないことから，影響が累積する線形閾値モデル（Linear Threshold Model）とは異なり，情報の内容を見た時点で行動を起こすかどうかが決定されるとする．もちろん，このアプローチでは，ユーザが情報を見ても行動を起こさない場合は考慮されないが，少なくともソーシャルメディアにおけるユーザの役目という観点から分析する場合は，そのようなユーザを無視したとしても分析結果には影響しない．

そうすることで，Twitter 上で観測できるユーザの行動から arXiv 論文 → 情報拡散者 → 情報収集者という関係を直接モデル化できると共に，情報流通経路の推定という誤差が生じる要素も排除できる．また，提案モデルにおける情報拡散者と情報収集者という異なる側面を同時に考慮すれば，ユーザの分類やユーザと bot の識別なども実現できると考えられる．

3.3 学術情報流通におけるユーザの特徴分析法

提案した 3 層モデルに基づいて，Twitter 上の学術情報流通という観点から，情報拡散者と情報収集者という異なる側面を持つユーザの特徴の分析を試みる．

まず，ユーザ $u_i (i = 1, \ldots, N)$ は，情報拡散者としての側面 u_i^s と情報収集者 u_i^c を持つとして，情報収集者 u_i^c が情報拡散者 u_j^s のツイートをリツイートやいいねした場合に，u_j^s から u_i^c に学術情報が伝播したと考え，その伝播の有無 $D_{i,j}$ を，式 1 のような N 行 N 列の隣接行列（adjacency matrix）$\boldsymbol{D}$ で表す．

$$\boldsymbol{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & r_{N,N} \end{bmatrix} \tag{1}$$

ここで，ソーシャルメディア上で有益と考えられる情報拡散者は多くの信頼できる情報収集者によってリツイートやいいねされるはずであり，その逆も成り立つと考えられることから，Kleinberg の **HITS**(Hyperlink-Induced Topic Search) アルゴリズム [10] を用いて，あるユーザの情報拡散者としての有益性と情報収集者としての信頼性を求める．

ユーザの情報拡散者としての重要度を表す HITS のオーソリティ度ベクトル $\boldsymbol{a} = (a_1, \ldots, a_N)^\top$，情報収集者としての重要度を表す HITS のハブ度ベクトル $\boldsymbol{h} = (h_1, \ldots, h_N)^\top$ は，隣接行列 $\boldsymbol{D}$ を用いた次の式の適用と正規化を収束するまで繰り返して求める．

$$\boldsymbol{a} = \boldsymbol{D}^\top \boldsymbol{h} \tag{2}$$

$$\boldsymbol{h} = \boldsymbol{D}\boldsymbol{a} \tag{3}$$

なお，各初期ベクトルの要素は 1 とする．

この結果，ユーザ u_i の特徴を，オーソリティ度とハブ度を組み合わせたベクトル (a_i, h_i) として定義する．さらにオーソリティ度とハブ度の順位を組み合わせて，各ユーザが学術情報流通という観点からどのような役目を果たしているかを分析する．

3.4 情報拡散者の関係ネットワークのクラスタ分析

学術情報流通において貢献度が高いオーソリティ度が高い情報拡散者に着目し，次の手順で情報収集者の観点から類似している情報拡散者のクラスタを抽出して，その特徴を分析する．

（1） オーソリティ度上位 M 人の情報拡散者 u_i^s を抽出する．

（2） 各情報拡散者 u_i^s にリツイート・いいねした情報収集者の集合 $U_i^c = \{u_j^c | d_{j,i} = 1\}$ を作成する．

（3） 情報拡散者 u_i^s と u_j^s の Jaccard 係数 $jaccard(U_i^c, U_j^c)$ が閾値 T 以上の場合にエッジ $e_{i,j}$ を張り，情報拡散者の関係ネットワーク G^s を作成する．

（4） Louvain 法 [11] で G^s を K 個のクラスタ $C_1^s, \ldots, C_K^s$ に分割する．

4 分　　析

4.1 データセット

本論文では，arXiv プレプリントサーバから OAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）[2]を用いて 2019 年 5 月 23 日に収集した 1,540,784 本の論文を収集し，arXiv データセットとして使用した．収集した論文情報には，論文 ID，著者名（複数可），投稿日，更新日，タイトル，カテゴリ（主・副あり，複数可），抄録，コメント，DOI，Journal reference，Report number，ACM class，MSC class が含まれる．論文は計算機科学（例，cs）などの 11 種類のアーカイブに収録され，カテゴリはアーカイブ名を大分類として，さらに詳細な分類の研究分野をピリオドで結合した文字列（例，cs.AI）である．

さらに，Twitter API を用いて 2011 年 1 月 2 日から 2019 年 5 月 15 日の間に arXiv 論文に関して発言したツイートを収集し，言及データセットとして使用した．発言の有無は，ツイート中の文字列”arxiv.org”と arXiv ID の有無で判定した．なお，arXiv ID の形式は表 1 に示すように 3 種類あるが，すべての形式に対応した．ただし，Twitter API にはツイートに関するリツイートといいねは最大 100 件までしか取得できない制限があったが，今回のデータセットには 100 件到達した事例はいいねは 1,492 件，リツイートは 432 件しかなかったので，分析の際に大きな問題にはならない．

4.2 基本統計量の分析

4.2.1 言及ツイートの分布

言及データの詳細を表 2 に示す．この結果から，arXiv 論文

2 : http://www.openarchives.org/OAI/openarchivesprotocol.html

表 1: arXiv ID の形式

時期	ID の形式の例
1991 年 7 月～2007 年 3 月	hep-th/9901001
2007 年 4 月～2014 年 12 月	0706.0001
2015 年 1 月～	1501.00001

表 2: Twitter の言及データセット

分析内容	値
ツイート総数	1,838,305
いいねされたツイート数	416,418
リツイートされたツイート数	245,119
言及されたユニーク論文数	763,735
言及したユニークユーザ数	75,708

に言及するユーザは限られること，学術情報が比較的高い確率でリツイート・いいねされること，いいねの方がリツイートよりも気軽に行われることなどがわかる．

さらに，論文の言及に関する特性を分析する．図 2a に横軸に言及論文数の順位，縦軸に言及人数を両対数を，図 2b に横軸に言及人数の順位，縦軸に言及した論文の数のプロットした両対数グラフを示す．図 2a は，1 位の論文だけ突出して言及数が多いことを除けば，一直線上のべき分布であることがわかる．また，図 2b は，中間は両端の直線上の分布のように見えるが，その中間では約 10^2 から約 10^4 への急激な増加が見られる．さらに，高順位ユーザの言及論文数は約 $10^4 \sim 10^6$ とかなり多いことから，高順位のユーザは bot である可能性が高いと思われる．

4.2.2 論文のカテゴリの言及割合の時系列変化の分析

Twitter 上でどのような分野の arXiv 論文が言及されているか，さらにその分野の時系列変化を分析した．論文の分野として，arXiv の主カテゴリだけを用いた．図 3 に，横軸に論文を投稿した年を，縦軸にその年の言及論文の主カテゴリの割合と言及数をそれぞれプロットしたグラフを示す．1 年単位でカウントしたので，分析対象期間は 2019 年を除く 2011～2018 年とした．なお，physics*は，”arXiv submission rate statistics”[3]の分類に合わせて，physics と gr-qc（General Relativity and Quantum Cosmology），nlin（Nonlinear Sciences），nucl（Nuclear Theory,Nuclear Experiment），quant-ph（Quantum Physics）を合わせたデータとした．この結果から，2011 年頃は math（数学）や physics*が中心であり，大部分の分野ではゆるやかな増加傾向を示しているのに対し，CS（計算機科学）だけは 2015 年から言及数と言及割合ともに急増し，2018 年の時点ではどちらもトップになっていることがわかる．

そこで，次に CS 分野に絞って，この時系列変化の違いが生まれる原因を，より詳細に分析した．図 4 に，横軸に論文を投稿した年を，縦軸にその年の CS 分野の言及論文の主カテゴリの割合と言及数をそれぞれプロットしたグラフを示す．この結果から，2011 年頃は cs.IT（情報理論）の言及割合が突出して多かったが，2015 年から特に cs.CV（画像認識）の言及数が急激な増加をはじめ，それに伴う cs.AI（人工知能）と cs.CL（自然言語処理）の言及数も大幅に増加しはじめていることがわかる．つまり，arXiv における CS 分野の論文言及の増加は，特に画像認識分野における深層学習の発展が原因と推測できる．

3 : https://arxiv.org/help/stats/2019_by_area/

表 3: ユーザの特性分類結果

分析内容	値
ユーザ数	293,969
authority > 0	38,476
hub > 0	279,718
authority , hub > 0	24,225
authority > 0 & hub = 0	14,251
hub > 0 & authority = 0	255,493

4.3 学術情報流通におけるユーザの特徴分析

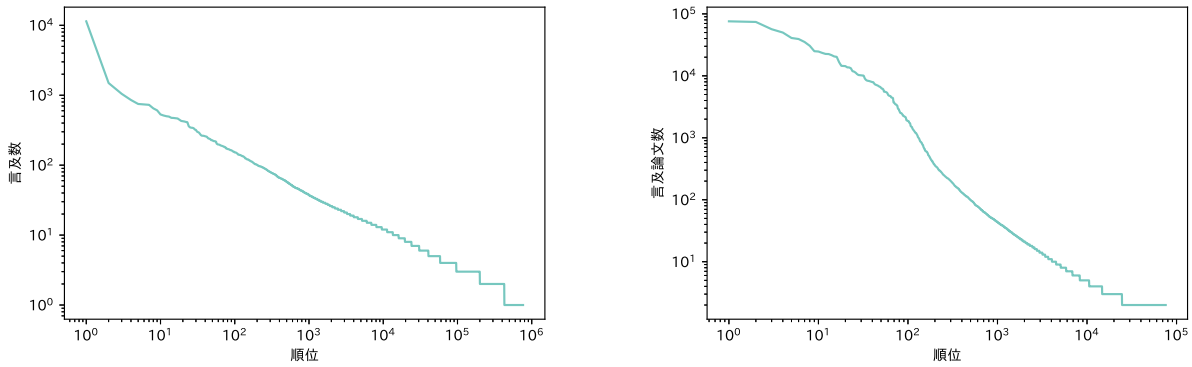
3.3 で述べた提案手法を用いて，Twitter 上の学術情報流通における情報拡散と情報収集という同一ユーザの異なる側面を分析した．

まず，図 5 に，各ユーザのオーソリティ度とハブ度の分布を示す．これを見ると，どちらも値が明確に大きいユーザはごく一部であることがわかる．特に，オーソリティ度が 0 付近に多くのユーザが集まっており，これだけでは詳しく分析することは難しい．

そこで，各値が 0 かどうかに着目して分類した結果を表 3 に示す．全ユーザ数は，言及・いいね・リツイートのいずれかを行なったユーザの総数であり，表 2 の言及ユーザ数の 3.88 倍であることから，Twitter 上で多くの人に影響を与えていると推測できる．しかし，オーソリティ度が 0 以上のユーザ数は言及ユーザ数の 50.1%であり，学術情報を拡散しても他の人に影響を与えない人も約半数いることがわかる．

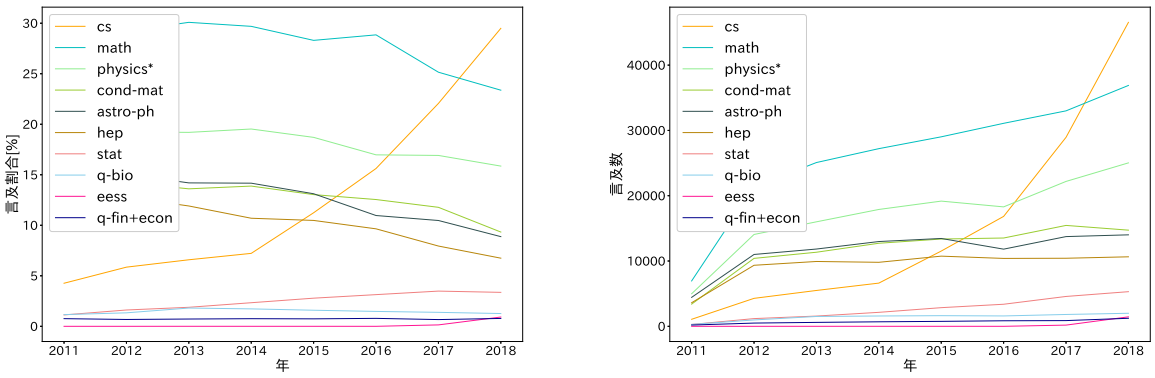
次にオーソリティ度とハブ度の組み合わせを調べると，オーソリティ度が 0 以上のユーザ数は，ハブ度が 0 以上のユーザ数の 13.8%であり，少ない限られた人数の情報拡散者が多くの情報収集者に影響を与えていることがわかる．そこで，オーソリティ度が 0 以上のユーザの内訳を調べると，ハブ度も正であるユーザは 58.8%であり，情報を拡散しても収集しないユーザが存在すると推測される．さらに，ハブ度が 0 以上のユーザの内訳を調べると，オーソリティ度が 0 であるユーザは 91.3%であり，単に情報を収集するだけのユーザが大部分であると言える．

そこで，実際にオーソリティ度・ハブ度が非常に高いユーザに関して，より詳しく分析する．オーソリティ度・ハブ度上位 20 人のユーザを，それぞれ表 4a と表 4b に示す．なお，括弧内は順位であるが，同値の場合でも同順位にはしていない．なお，ユーザ名が欠落しているユーザが存在するが，これは言及ツイートがないユーザに対してはユーザプロファイルを収集しなかったからである．表 4a を見ると，表 3 のオーソリティ度が高くてもハブ度が 0 のユーザは，arxiv，Statistics Papers, arXiv CS-CV のような bot であることがわかった．つまり，学術情報流通に限れば bot が有益な役割を果たしていることが確認できる．また，オーソリティ度の上位には国際的に著名な研究者の他に，DeepMind といった有名な企業ユーザも現れている．さらに，図 2b の 2 種類の分布と照合することで，新着論文をそのまま言及する bot 群と，論文を選別して言及する人間という 2 種類のユーザの存在により生まれたと推測できる．表 4b を見ると，オーソリティ度の順位が 1 桁のユーザが 1 人，3 桁の



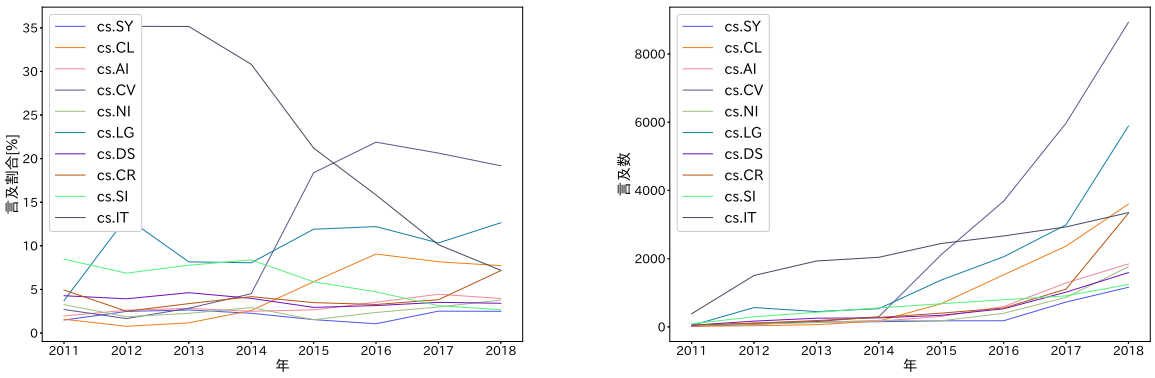
a 論文の言及人数の分布　　b ユーザの言及論文数の分布

図 2: 言及人数とユーザの言及論文数の分布



a 言及割合　　b 言及数

図 3: 言及論文の主カテゴリの時系列変化



a 言及割合　　b 言及数

図 4: CS 分野の言及論文の主カテゴリの時系列変化

ユーザが 4 人存在するが，同時にオーソリティ度が 0 のユーザが 5 人存在することがわかる．つまり，高いオーソリティ度と低いオーソリティ度のユーザが混在しているが，これは情報拡散にはまったく関与せずに，情報収集だけを非常に積極的に行うユーザが存在することを示している．

4.4 情報拡散者の関係の分析

次に，学術情報流通に対する貢献度が高い高オーソリティ度の情報拡散者同士の関係について分析する．

オーソリティ度の上位 1000 人を抽出し，で述べた提案手法を用いて，作成した関係ネットワークを複数のクラスタに分割

表 4: 上位 20 人のユーザ

a オーソリティ度

	ユーザ名	オーソリティ度	ハブ度
1	Miles Brundage	0.013766	0.000262 (93)
2	arxiv	0.012126	0.000000 (283715)
3	Alex J. Champandard	0.009185	0.000140 (620)
4	hardmaru	0.007347	0.000382 (19)
5	samim	0.007254	0.000197 (268)
6	Ian Goodfellow	0.007096	0.000120 (839)
7	François Chollet	0.006511	0.000127 (755)
8	Andrej Karpathy	0.006438	0.000103 (1130)
9	Nenad Tomasev	0.006435	0.000287 (64)
10	Tomasz Malisiewicz	0.006300	0.000118 (871)
11	DeepMind	0.005748	0.000005 (38650)
12	Statistics Papers	0.005508	0.000000 (289968)
13	Nando de Freitas	0.005137	0.000174 (369)
14	Eclipse DL4J	0.004999	0.000175 (364)
15	fastml extra	0.004979	0.000202 (244)
16	Denny Britz	0.004885	0.000104 (1114)
17	John Platt	0.004528	0.000040 (4663)
18	arXiv CS-CV	0.004368	0.000000 (283147)
19	Thomas Lahore	0.003970	0.000248 (127)
20	Soumith Chintala	0.003736	0.000107 (1063)

b ハブ度

	ユーザ名	オーソリティ度	ハブ度
1		0.000000 (252352)	0.000599
2		0.000000 (289690)	0.000530
3	fly51fly	0.000071 (1915)	0.000501
4	Igor Carron	0.000105 (1429)	0.000467
5	priya joseph	0.000000 (40817)	0.000458
6	Emmanuel Kahembwe	0.000140 (1109)	0.000456
7	闇堕ちくん	0.000074 (1856)	0.000451
8	Ben Duffy	0.000111 (1368)	0.000430
9	Atul Acharya	0.000031 (3629)	0.000419
10	Robert Dionne	0.000192 (847)	0.000407
11	Hamid	0.000209 (778)	0.000407
12		0.000000 (241748)	0.000404
13	Montreal.AI	0.000723 (210)	0.000398
14		0.000000 (200803)	0.000396
15	Richard Kelley	0.000039 (2996)	0.000394
16	Vincent Boucher ?	0.000008 (10881)	0.000390
17	mat	0.000690 (223)	0.000390
18	Ed Henry	0.000085 (1670)	0.000383
19	hardmaru	0.007347 (4)	0.000382
20		0.000000 (273095)	0.000378

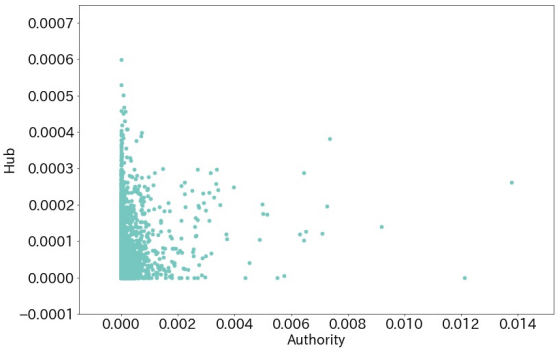


図 5: オーソリティ度とハブ度の分布

した結果，18 個のクラスタが得られた．なお，閾値 T は，2 つのノードの Jaccard 係数とその順位をプロットした時に，急激に値が低下する直前の 0.15 とした．

関係ネットワーク全体の可視化結果の連結成分の一部を，図 6 に示す．関係ネットワークは 13 個の連結成分で構成され，それが提案手法で 18 個のクラスタに分類されるが，サイズが 4 以上の連結成分は 5 個だけだけであり，その中から特徴的な 3 個の連結成分を選んで Cytoscape で可視化した．ノードの色はクラスタ別に彩色し，ノードの直径はオーソリティ度に従って大きくなるように，エッジの濃度と太さは Jaccard 係数に従って大きくなるように描画した．

図 6a に示す連結成分 1 は，ノード数 48 と最大であり，5 個のクラスタを含んでいた．ユーザ名が日本語やそのローマ字表記が多く，さらに Daisuke Okanohara や，Hideki Nakayama などの機械学習分野の産学の著名な日本人研究者や，Graham Neubig や Danishka Bollegala などの日本の大学に現在在籍している，あるいは在籍していた研究者が見られることから，この連結成分は日本の機械学習の研究者達であることがわかる．

図 6b に示す連結成分 2 は，ノード数 29 であり，2 個のクラスタを含んでいた．ユーザ名を手掛かりに調べると，OpenAI の Miles Brundage や Andrej Karpathy，GANs の Ian Goodfellow，PyTorch の開発者である Soumith Chintala，Google Brain の Dustin Tran などの著名な機械学習分野の研究者が多いことから，国際的な機械学習の研究者達であることがわかる．

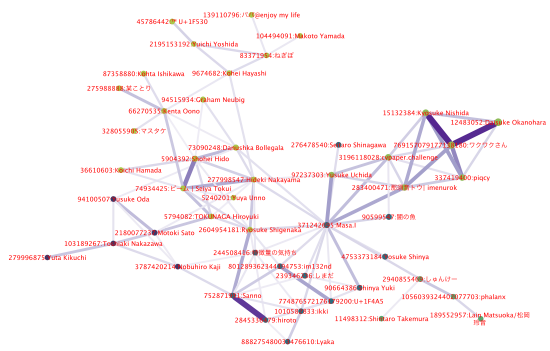
図 6c に示す連結成分 3 は，ノード数 4 と小さく，1 個のクラスタを含む．ユーザ名は，math.CT, math.AT, math-ph, hep-th などの数学や数理物理学の arXiv の bot であることがわかる．図 3 を見る限りでは計算機科学系と大きく変わらないことから，これらの分野の研究者は Twitter を主に arXiv の bot から直接学術情報を得る手段として使用していると推測され，研究者間のコミュニケーションは学会や打ち合わせにおける対面の議論など，別の手段で実行している可能性が高い．

4.5 情報拡散者クラスタの特徴分析

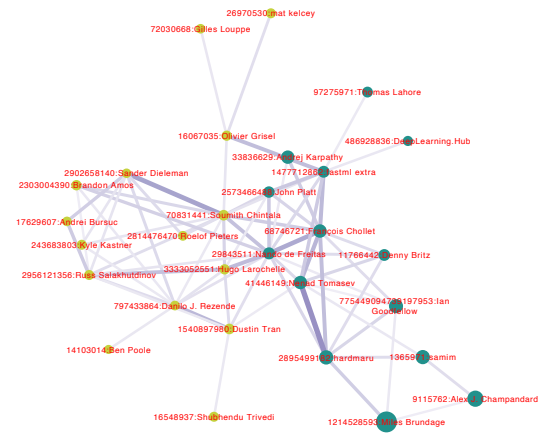
最後に，情報拡散者のクラスタの情報拡散者・情報収集者としての特性を，arXiv 論文のカテゴリを用いて分析する．表 5 に，各クラスタが属している連結成分と，そのクラスタによって言及されたカテゴリを示す．実際には，ユーザ単位で arXiv 論文の主カテゴリの頻度を集計し，最も頻度が高いカテゴリをユーザの専門性を反映しているカテゴリとみなして，さらに各カテゴリごとにユーザ数を集計して，ユーザ数が多い順に左から並べた．なお，クラスタ 8 は情報を拡散しただけでなので，情報収集者としてのカテゴリは取得できなかった．

この結果から，図 6a と図 6b の連結成分 1 と 2 の情報拡散者は，cs.LG（機械学習）, cs.CV（画像認識）とほとんど同じ分野であることが確認できる．一番サイズが大きい連結成分 1 の比較的明確の違いは，クラスタ 3, 4, 13 などの複数のクラスタに cs.CL（自然言語処理）が含まれていること，クラスタ 3 の情報拡散者と情報収集者の両方に stat.ML（機械学習）と cs.DS（データ構造とアルゴリズム）が含まれていることである．また，クラスタ 2, 9, 15, 16 では，情報拡散者の分野と異なる cs.LG が情報収集者として抽出されており，昨今の機械学習の発展は多彩な分野の研究者の関心を引いていることがわかる．

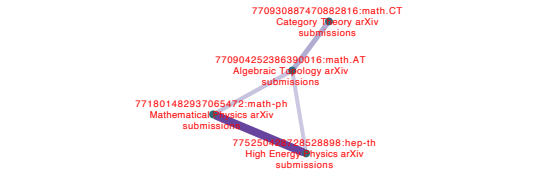
興味深い点は，連結成分 2（クラスタ 1, 5）には機械学習分野を牽引している世界的に著名研究者が多く，さらにツイート



a 連結成分 1



b 連結成分 2



c 連結成分 3

図 6: オーソリティコミュニティの可視化結果

情報収集時に日本からのツイートを優先するようなバイアスを一切掛けていないにもかかわらず，連結成分 1 の日本の機械学習の研究者の集団が多く抽出されていることである．

そこで，連結成分 1 の日本の機械学習研究者のグループ（日本グループ）と，連結成分 2 の国際的な機械学習研究者（国際グループ）のグループの特徴の違いを分析する．

まず，両者の言及した論文集合について調べると，日本グループは 2447 本，国際グループは 9333 本であり，共通論文は 1101 本であった．日本グループは独自の論文も多いことから，単なる海外グループが拡散した情報の翻訳や要約だけではなく，独自の情報を発信している集団である可能性が考えられる．

そこで各グループが拡散する情報の詳細の違いを調べるために，今度は言及した論文のすべてのカテゴリを集計し，図 7 に出現頻度上位 10 件のカテゴリを頻度順に並べて分析する．図 7a と図 7b を比較すると，上位 3 位までは cs.CV, cs.LG, stat.ML と一致し，4, 5 位は cs.CL と cs.AI の順位が逆転しているが，6 位は cs.NE（ニューラルネットと進化計算）で再び一致することから，興味の種類の傾向はほぼ一致していることがわかる．ただし，出現頻度を見ると，日本グループでは上位 3 位までが極めて高く，特にこの分野の研究者が Twitter 上の学術情報流通に大きな影響を与えていると考えらる．これに対して，国際グループでは頻度は緩やかに減少しており，cs.RO（ロボット工学）や cs.IR（情報検索）の頻度も相対的に高いと言える．

4.6 考　　察

今までの結果についてまとめて考察する．

まず，学術情報流通においては，研究者だけではなく，arXiv などの公式 bot もかなりの割合で有用な役目を果たしていることがわかった．従来のソーシャルメディア分析では bot は分析対象から除外することが多かったが，学術情報流通では bot を許容した分析手段を確立する必要がある．

次に，深層学習の急速な発展は，arXiv におけるプレプリントの閲覧と，Twitter 上の学術情報流通に支えられていることが明らかになった．特に後者は arXiv のプレプリントを活用している他の分野には見られないほど顕著なものであり，論文著者や関係者による直接の情報発信や，研究者同士の議論が行われていると考えられる．

さらに，機械学習分野の学術情報の拡散においては，国際的な著名研究者のグループとは独立して，日本という限定された地域のグループが存在し，しかもその方が優位であることがわかった．さらに，この国際グループと日本グループにおけるカテゴリ分布の差がほとんどないことから，局地的に独自の発展をしているのではなく国際的な動きと連動していることがわかったが，言及論文の重なりを考えると，海外の情報をそのまま翻訳・要約しているわけではなく，活動の独自性もあることがわかった．このような差が生じた理由としては，一般的に Twitter の利用度が高い日本では学術情報流通においても他の国より積極的に活用されていること，日本語はインド・ヨーロッパ語族である英語は独立に発展したために日本語での学術情報を望むユーザが多いこと，日本は国際的な学術コミュニティとは多少距離があることから cvpaper.challenge[4]のように Twitter 上で連携してトップレベル国際会議を目指したり，国際会議動向を広く伝えようとする研究者が多いこと，などが考えられるが，その検証のためには，さらなる分析が必要である．

5　おわりに

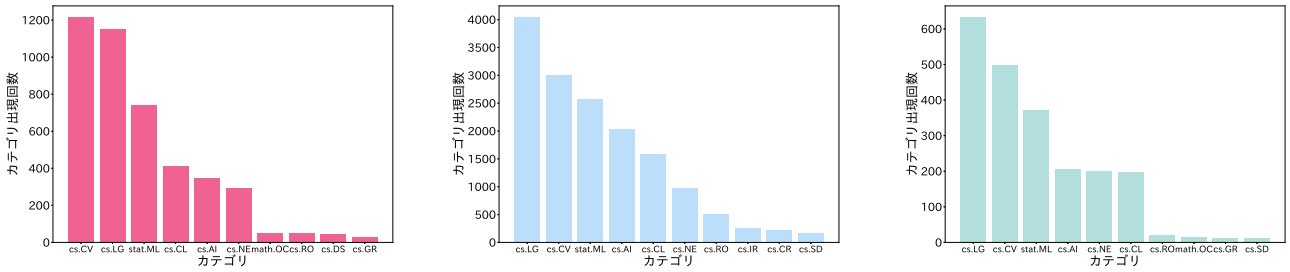
本論文では Twitter 上のユーザには学術情報の流通と獲得の 2 種類の側面があるとして，彼らの行動からその相補的な関係を 3 層モデルに基づいて分析した．その結果，arXiv の bot や Twitter が arXiv 論文情報流通において重要な役目を果たしていることを明らかにすると共に，高オーソリティ度の情報拡散者の関係ネットワークから，近年急速に発展している機械学習関係に関して積極的な貢献をしている国際グループと日本グループの存在を確認できた．

また，学術情報流通に対する貢献度が高い高オーソリティ度

4 : https://twitter.com/cvpaperchalleng

表 5: クラスタの特徴

連結成分	クラスタ	情報拡散者	情報収集者
1	3	stat.ML,cs.DS,cs.CV,cs.LG,cs.CL	stat.ML,cs.DS,cs.CV,cs.LG,math.ST,cond-mat.stat-meth
	4	cs.LG,cs.CL,cs.CV	cs.CV,cs.LG
	6	cs.LG,cs.CV	cs.LG,cs.CV
	10	cs.LG,cs.CV	cs.LG,cs.CV
	13	cs.CL	cs.LG,cs.CL
2	1	cs.LG,cs.CV	cs.LG,cs.CV
	5	cs.LG,cs.CV,stat.ML	cs.LG,cs.CV
3	14	math.AT,math.CT,math-ph,hep-th	math.AT,math.CT,math-ph,hep-th
4	9	cs.CL	cs.LG,cs.CL
5	0	hep-th	quant-ph,cs.LG,hep-th,math-ph
6	7	physics.data-an,physics.soc-ph	physics.soc-ph
7	11	cond-mat.stat-mesh	cond-mat.stat-mesh
8	17	q-bio.NC	physics.soc-ph,q-bio.NC
9	2	cs.IT,cs.LG	stat.ML,cs.LG
10	8	hep-th,math-ph	
11	12	cond-mat.str-el	quant-ph
12	15	cs.SD,eess.AS	cs.LG,cs.CL
13	16	hep-th,quant-ph	cs.LG,cs.CL



a 日本グループ　　b 国際グループ　　c 両方の共通集合

図 7: 日本・国際グループで言及されている論文のカテゴリ分布

の情報拡散者同士から関係ネットワークを得た結果，CS 分野，特に機械学習分野のグループとして国際グループと日本グループが得られており，さらに，ヨーロッパ圏，アジア圏などで他のローカルなグループはないことから，なぜ日本グループだけが国際グループとは別に存在するかという理由と活動の違いについて，共通で言及されている arXiv 論文の言及タイミングの違いなどを用いて分析する予定である．

さらに，ソーシャルメディアにおける学術情報流通には分野によるアクティビティの違いが顕著であることが判明したことから，分野特性を考慮した arXiv 論文の推薦システムやオルトメトリクスのような評価指標を検討する予定である．

謝　　辞

本研究は JSPS 科研費 19H04421 の助成を受けて行った．

文　　献

[1] Vincent Larivière, Stefanie Haustein, and Philippe Mongeon. The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE*, Vol. 10, No. 6, pp. 1–15, 06 2015.

[2] Andrea Chiarelli, Rob Johnson, Stephen Pinfield, and Emma Richens. Accelerating scholarly communication: The transformative role of preprints, September 2019.

[3] Qing Ke, Yong-Yeol Ahn, and Cassidy R. Sugimoto. A systematic identification and analysis of scientists on Twitter. *PLOS ONE*, Vol. 12, No. 4, pp. 1–17, 04 2017.

[4] Ehsan Mohammadi, Mike Thelwall, Mary Kwasny, and Kristi L. Holmes. Academic information on Twitter: A user survey. *PLOS ONE*, Vol. 13, No. 5, pp. 1–18, 05 2018.

[5] Vincent Larivière, Cassidy R. Sugimoto, Benoit Macaluso, Staša Milojević, Blaise Cronin, and Mike Thelwall. arxiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, Vol. 65, No. 6, pp. 1157–1169, 2014.

[6] Xin Shuai, Alberto Pepe, and Johan Bollen. How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PLOS ONE*, Vol. 7, No. 11, pp. 1–8, 11 2012.

[7] Dennis Kergl, Robert Roedler, and Sebastian Seeber. On the Endogenesis of Twitter's Spritzer and Gardenhose Sample Streams. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '14, p. 357–364. IEEE Press, 2014.

[8] Gunther Eysenbach. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *J Med Internet Res*, Vol. 13, No. 4, p. e123, Dec 2011.

[9] Cassidy Sugimoto, Benot Macaluso, Timothy Bowman, Stefanie Haustein, Vincent Larivire, and Katy Brner. Measuring Twitter activity of articleXiv e-prints and published papers. *figshare*, 01 2014.

[10] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, Vol. 46, No. 5, pp. 604–632, September 1999.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008, oct 2008.