ソーシャルネットワークの特徴に基づいた有向ネットワーク生成モデルの提案

山川　衛[†]　田島　敬史[†]

† 京都大学情報学研究科　〒 606-8501 京都府京都市左京区吉田本町 36-1
E-mail: †yamakawa.mamoru.75s@st.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　Web グラフやソーシャルネットワークサービス (SNS) 上のフォローグラフにおいては，相互フォローが重要な役割を果たしていることが多く，近年はこれらに着目したネットワークの分析に関する研究が盛んになっている．また，近年はクラスター係数を高めるために Edge Rewiring を採用したネットワーク生成モデルも多数存在するが，非常に実行時間がかかるため，Edge Rewiring を用いない同程度の特徴を持ったモデルの提案は重要であると考える．そこで本稿では，まず，日本の Twitter ユーザのフォローネットワークを解析し，その結果にもとづき相互フォローを考慮した新たな枝の選択方法を用いる有向ネットワーク生成モデルを提案する．また，このモデルにより生成したネットワークと実際の SNS との性質の類似性を比較検証する．

キーワード　ソーシャルネットワークグラフ，ランダムグラフ生成，成長ネットワーク

1 はじめに

1998 年に Watts–Strogatz (WS) モデル [1] が提案されて以来，様々な人工的なネットワーク生成手法が提案されてきている．現在は Twitter をはじめとした SNS 上のフォローグラフを研究で扱うことも多くなってきており，相互エッジが存在する有向ネットワークの効率的な生成は重要であると考える．そこで本稿では Twitter のフォローグラフなどの相互エッジが存在する有向ネットワークを生成する手法を提案する．

今回は日本の Twitter ユーザ 27,500,000 人によるネットワークを解析し，現在の Twitter のネットワークがどのような性質を持っているかの解析を行った．その結果，ネットワークの次数分布は低次数ではべき乗分布に沿わないことが判明した．

この結果をもとに，本稿ではユーザの優先的選択を考慮した有向ネットワーク生成手法を提案した．提案手法はノードが新規接続を行う際に優先的選択の方法を複数から選択するものとなっている．パラメータの変化による次数分布の変化の観察を行った結果，提案モデルはパラメータの調整によって実データの低次数で見られる分布と高次数で見られる従来のべき乗分布の両方を表現できることが判明した．この結果から，実際の SNS ではノードの次数によってパラメータが変化していると考えられるため，今後は次数によってパラメータを変化させるモデルを作成し検証していきたい．

2 関連研究

ここではネットワーク生成モデルの中でも，成長ネットワークに関する先行研究を説明する．

2.1 BA モデル

Barabási-Albert モデル [2] は m_0 個の頂点から開始し，目標の頂点数になるまで以下を繰り返すモデルである．

（1） 新しい頂点をエッジを m 個持った状態でネットワークに追加する

（2） 次数に応じた確率で m 個の頂点を選択し，新しい頂点が持つエッジを接続する．

BA モデルの次数分布はべき乗測に従い，スケールフリー性を持つ．

2.2 Holme-Kim モデル

Holme-Kim モデル [3] は，m_0 個の頂点から開始し，目標の頂点数になるまで以下を繰り返すモデルである．

（1） 新しい頂点をエッジを m 個持った状態でネットワークに追加する

（2） 次数に応じた確率で頂点 v を選択し，新しい頂点が持つエッジを接続する．

（3） 確率 p で v の隣接点を 1 つ選び，新しい頂点が持つエッジを接続する．確率 $1-p$ で (2) と同様の優先的選択を行う．

（4） 残り $m-2$ 本のエッジの接続先を (3) と同様に決定する．

Holme-Kim モデルは BA モデルの拡張の一つであり，BA モデルのクラスター係数が小さいという欠点を克服するために提案されため，スケールフリー性と高クラスター性を持つ．

2.3 CNN モデル

Connecting nearest-neighbor (CNN) モデル [4] は 1 つの頂点と空の辺集合から開始し，目標の頂点数になるまで以下を繰り返すモデルである．

（1） 確率 $1-\mu$ で新しい頂点をグラフに追加し，ランダムに頂点 j を選択し新規頂点と接続する (このとき新規頂点と j のすべての neighbor との間に潜在エッジを張る)

（2） 確率 μ で潜在エッジをランダムに選択し，エッジに変

表 1　データセットの各指標

$\|V\|$	27,500,000
$\|E\|$	8,740,471,435
reciprocal follow	996,963,609
one-way follow	6,746,544,217
D	2.05×10^{-5}

換する．

BA モデルは優先的選択を行うがクラスター性はなく，CNN モデルは優先的選択を行わないがスケールフリー性とクラスター性を表現できるモデルとなっている．

3　データセット

3.1　概　　要

今回は，2015 年に収集した lang=ja の約 4 千万の Twitter ユーザのうち，その当時のフォロワー数が多い方から 2750 万ユーザについて，2021/12（から 2022/1 にかけて）時点でのフォロワー，フレンドのリストを Twitter 社から購入して作ったフォローネットワークを解析した．フォローネットワークは有向グラフ $G = (V, E)$ で定義され，各 $i \in V$ はユーザを，$(i, j) \in E$ は i が j をフォローしているという関係を表す．なお，枝の両端のうちどちらかが 2750 万ユーザに含まれてない枝はフォローネットワークの作成時に削除している．また，ネットワークを解析する際の指標として以下を定義する．

In-degree

ユーザ i の In-degree $I(i)$ は i の入次数であり，これは Twitter におけるフォロワー数に該当する．

Out-degree

ユーザ i の Out-degree $O(i)$ は i の出次数であり，これは Twitter におけるフレンド数に該当する．

Reciprocal-degree

ユーザ i の Reciprocal-degree $R(i)$ は $(i, j), (j, i) \in E$ となるような $j \in V$ の数で定義する．これは Twitter における相互フォロワー数に該当する．

Degree

ユーザ i の次数 $\deg(i)$ は $I(i) + O(i)$ で定義する．

Reciprocal follow

Reciprocal follow は $\frac{1}{2}\sum_i R(i)$ で定義する．

One-way follow

One-way follow は $|E| - \sum_i R(i)$ で定義する．

密度

有向グラフの密度 D は $\frac{|E|}{|V|(|V|-1)}$ で定義される．

3.2　解 析 結 果

27,500,000 ノード ($|V| = 27,500,000$) 全体が作るネットワークにおいて，各指標は図 1 のようになった．図 1，2 はそれぞれデータセットにおける Reciprocal-degree, In-degree, Out-degree, degree の確率分布と累積分布関数をプロットしたものである．図 1 から各次数分布は低次数帯ではべき乗則に従っておらず，

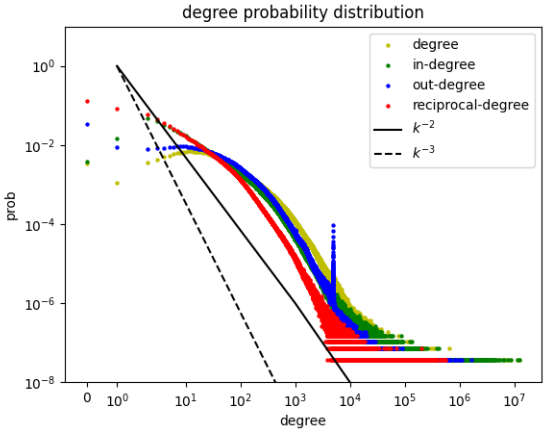

図 1　次数の確率分布

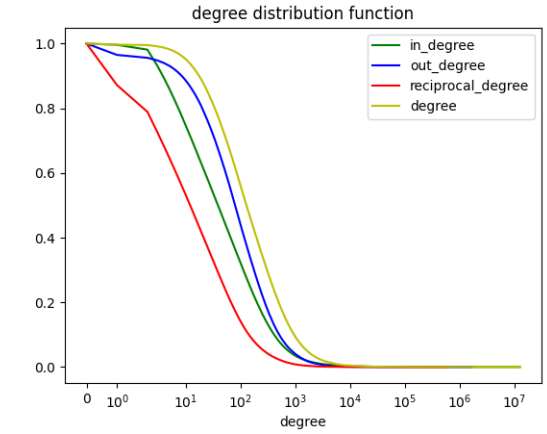

図 2　次数の確率の累積分布関数

山なりの分布となっており，高次数では本来のスケールフリーネットワークにみられるロングテールの分布となっていることが分かる．また，Out-degree が 10^3 と 10^4 の間に集中しているが，これは Twitter のフォロー制限の上限までフォローを行っているユーザが多いためと考えられる．

4　モデル概要

この節では，提案手法のモデルの解説を行う．なお，提案モデルの疑似コードは Algorithm1，2 に記載している．

提案モデルは入力として頂点数 N とパラメータ μ_1，μ_2，α が与えられ，ネットワークの頂点数が N になるまで確率的に以下を繰り返すモデルである．

（1）　確率 μ_1 でフォロー行動を行う既存ノード i をランダムに選択する．その後，i は入次数に応じた確率でノード j をランダムに選択し接続を行う．その後，j は確率 α で i と接続を行う．

（2）　確率 μ_2 でフォロー行動を行う既存ノード i をランダムに選択する．その後，i は $(A^2)_{ij}$ に応じた確率でノード j をランダムに選択し接続を行う．その後，j は確率 α で i と接続を行う．

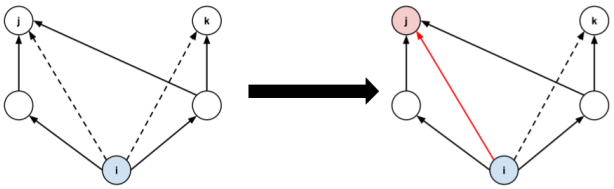

図 3　2-hop 先の優先的選択

（3） 確率 $1-(\mu_1+\mu_2)$ で新規ノード i をネットワークに加える．この時，新規ノードは入次数に応じた確率で既存ノード j を選択し接続する．その後，j は確率 α で i と接続を行う．

ここで，A は各試行時点でのネットワーク全体の隣接行列で，A_{ij} は枝 (i,j) が存在する場合に 1，存在しない場合には 0 となる行列であり，μ_1，μ_2，α は以下の条件を満たす実数である．

- $0 \leqq \mu_1+\mu_2 < 1$
- $\mu_1 \geqq 0$
- $\mu_2 \geqq 0$
- $0 \leqq \alpha \leqq 1$

また，(1)，(3) でノード i が接続先としてノード j を選択する確率，(2) でノード i が接続先としてノード j を選択する確率はそれぞれ以下の値となる．

- $\dfrac{I(j)+1}{\sum_{j\neq i}(I(j)+1)}$
- $\dfrac{(A^2)_{ij}}{\sum_{j\neq i}(A^2)_{ij}}$

$(A^2)_{ij}$ は i から j までの 2-hop の経路の数，つまり j をフォローしている自分のフレンドの数となる．例えば図 3 では i のフォロー候補としては j と k が存在するが，$(A^2)_{ij}=2$，$(A^2)_{ik}=1$ であり，j は k よりも 2 倍の確率で優先的に選択される．

(1)，(3) における優先的選択は BA モデルにおいてノードが新規参入した際の優先的選択行動と同一の行動であり，また (2) における 2-hop 先の優先的選択は Twitter では「自身のフレンドの多くがフォローしているフレンドをフォローする行為」に該当し，Twitter 上でも起こる可能性の高いフォロー行動であり，i が $(A^2)_{ij}$ の値が大きい j をフォローした場合の方がネットワークに含まれる三角形の数が増加するため，Holme Kim モデルや CNN モデルと同様にクラスター性を高めるための行動である．各試行におけるユーザの行動は Holme Kim モデルと近いものとなっているが，Holme Kim モデルでは 2-hop 先のノードを選択する際は最初に接続した頂点の隣接点しか選択できないのに対して提案モデルでは全ての 2-hop 先のノードを選択できる点，提案モデルは 2-hop 先のノードも優先的選択を行うという点で異なっており，また確率 α による接続は Twitter におけるフォローバックに該当し，フォローバックを明示的に考慮する点も本モデルの特徴となっている．

5　実験・評価

5.1　μ_1, μ_2 の比率による変化

ここでは提案モデルによって生成されたネットワークと実

Algorithm 1 generate_graph(N, μ_1, μ_2, α)

Require: $N \in \mathbf{N}, \mu_1, \mu_2 \geqq 0, 0 \leqq \mu_1+\mu_2 < 1, 0 \leqq \alpha \leqq 1$

$V = \{0\}$
$E = \emptyset$
$G = (V, E)$
$n = 1$
while $n < N$ **do**
  p = rand()
  A = adjacency matrix of G
  if $p \leqq \mu_1$ **then**
    randomly select i from V
    select j with probability $\frac{I(j)+1}{\sum_{j\neq i}(I(j)+1)}$
  else if $p \leqq \mu_1+\mu_2$ **then**
    randomly select i from V
    select j with probability $\frac{(A^2)_{ij}}{\sum_{j\neq i}(A^2)_{ij}}$
  else
    $i = n$
    randomly select j from V
    $V = V \cup i$
    $n = n+1$
  end if
  add_edge(G, i, j, α)
end while

Algorithm 2 add_edge(G, i, j, α)

Require: Directed graph $G = (V, E), i, j \in V, 0 \leqq \alpha \leqq 1$

if $(i, j) \notin E$ **then**
  $E = E \cup (i, j)$
  p = rand()
  if $p \leqq \alpha$ and $(i, j) \notin E$ **then**
    $E = E \cup (j, i)$
  end if
end if

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
$\lvert V\rvert$	10,000						
μ_1	0	0.15	0.30	0.45	0.60	0.75	0.9
μ_2	0.9	0.75	0.60	0.45	0.30	0.15	0
α	0.1						

表 2　各グラフの生成時のパラメータ

データとの比較，また三つの行動の比率によるネットワークの変化の観察を行う．今回は表 2 で示すように頂点数 $|V| = 10,000$，$\alpha = 0.1$ を固定し，μ_1, μ_2 を変化させた 6 グラフのネットワーク指標を分析した．各グラフについて各指標は表 3 のように，各グラフ次数分布と累積分布関数はそれぞれ図 4 から 17 のようになった．

図 4 から 17 から見て取れるように，reciprocal-degree をのぞく三つの次数分布がパラメータ μ_2 の割合が増加するほど山なりになっていき，テールの部分が短くなっている．この山なりの分布は図 1 の低次数で得られた分布の傾向と一致しており，低次数のユーザのフォロー行動の多くは提案モデルの (2) に示す 2-hop 先のユーザに接続する行動である可能性がある．実データは図 1 で見られるように低次数ではこの山なりの分布，高次

表 3　各ネットワークの指標の値

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
$\|V\|$	10,000						
$\|E\|$	109,268	108,030	109,067	109,637	108,022	107,657	103,137
reciprocal follow	10,446	10,196	10,286	10,386	10,460	10,592	10,289
one-way follow	88,376	87,637	88,495	88,864	87,101	86,473	82,559
D	0.00109	0.00108	0.00109	0.00110	0.00108	0.00108	0.0103
$\langle C \rangle$	0.00165	0.00190	0.00257	0.00487	0.01238	0.03321	0.12241

表 4　各ネットワークの指標の値

	G'_1	G'_2	G'_3	G'_4	G'_5	G'_6	G'_7
$\|V\|$	10,000						
$\|E\|$	140,955	136,350	134,705	129,837	124,622	117,884	104,949
reciprocal follow	42,804	39,504	36,854	32,598	27,871	21,844	13,198
one-way follow	55,346	57,342	60,997	64,641	68,880	74,195	78,552
D	0.00141	0.00136	0.00135	0.00129	0.00125	0.00118	0.00105
$\langle C \rangle$	0.00857	0.00922	0.01181	0.01914	0.03081	0.06749	0.17760

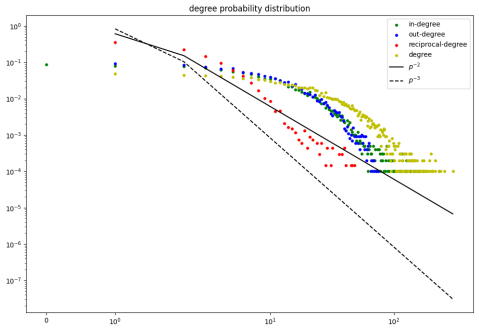

図 4　G_1 の次数分布

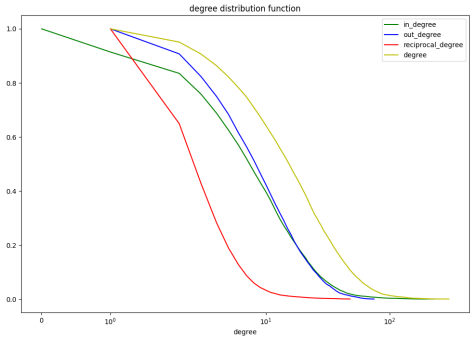

図 5　G_1 の累積分布関数

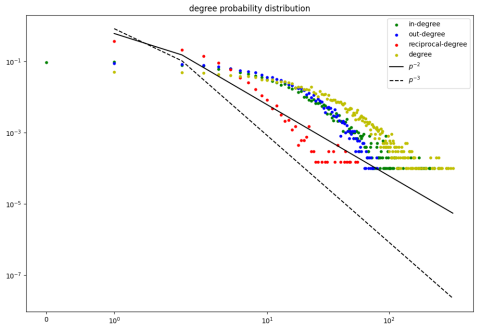

図 6　G_2 の次数分布

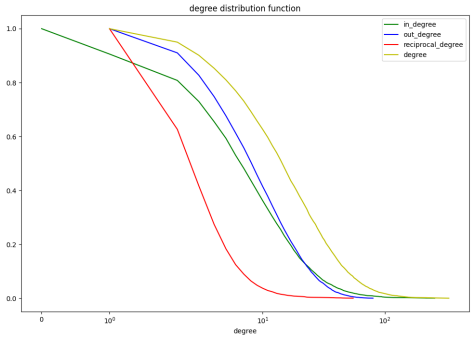

図 7　G_2 の累積分布関数

数ではロングテールをもつべき乗分布となっているため，次数に応じて μ_1, μ_2 の値が変わることによって次数分布を表現できる可能性がある．また reciproccal-degree の分布についても同等の変化がみられるがパラメータの変化による影響がその他の次数分布より少なく，この実験で変化をさせていない α によって分布が大きく変化すると考えられる．

5.2　α の設定方法による変化

先ほどのモデルではフォローバックの際にノードの特徴に関係なく一定の確率を用いていたが，2 ノードの次数の積に応じて接続確率を決定する Chung-Lu モデル [5] のように，次数に応じて 2 点の接続確率を決定するものも存在する．そのためここでは，α を変化させることで発生する分布の変化を観察する．本節では α のかわりに提案モデルの各試行において j が i に接続する確率 α_{ji} を以下で定義する．

$$\alpha_{ji} = \frac{I(i) + 1}{I(i) + I(j) + 2}$$

α_{ji} は $I(i)$ よりも $I(j)$ が大きければ大きくなり，j は自分より In-degree の大きい i ほど接続 (フォローバック) しやすくなる．この α_{ji} は，より Twitter 上のフォローバックに近い確率でフォローバックが起こると考える．

α_{ji} が与える影響を調べるため，α 以外のパラメータを表 2 と同様に設定し，パラメータ α として α_{ji} を採用したグラフ G'_1 ～G'_7 を作成し，グラフの性質の変化を観察した．生成した G'_1 ～G'_7 について，G_1～G_7 と同様に解析した結果，各指標は表 4，次数分布が図 18～図 31 になった．

各 G_i と G'_i を比較すると，reciprocal-degree 以外の分布は同

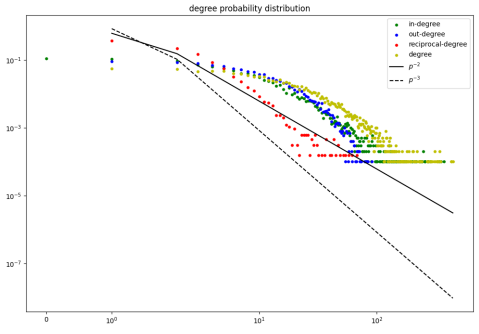

図 8　G_3 の次数分布

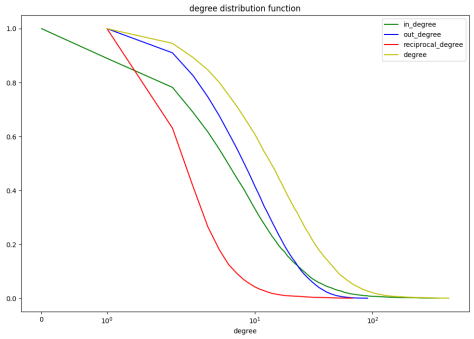

図 9　G_3 の累積分布関数

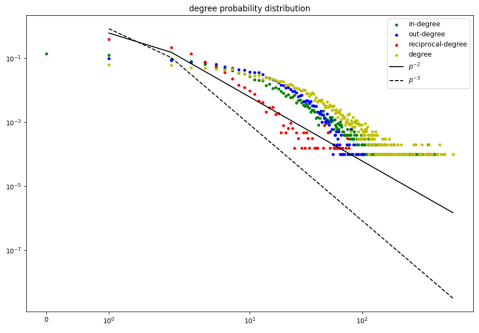

図 10　G_4 の次数分布

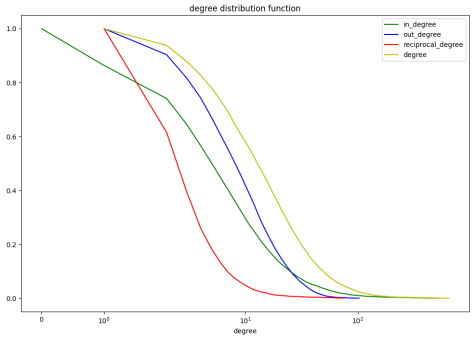

図 11　G_4 の累積分布関数

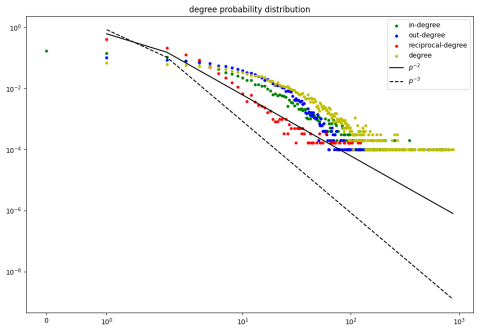

図 12　G_5 の次数分布

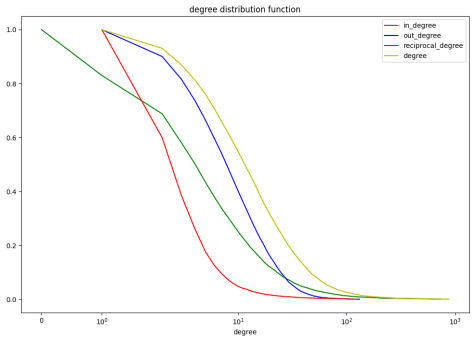

図 13　G_5 の累積分布関数

様の性質を見せているが reciprocal-degree の分布に変化がみられ，G'_i はよりその他の分布に形がより近づいている．これは α_{ji} を用いたことで $I(i)$ に応じて $R(i)$ が増えやすくなったためであると考えられる．また，μ_1 が上昇するほど reciprocal follow の割合が減少するという性質がみられた．i が j をフォローした場合，$I(j)$ は 1 増加するが $\sum_k (A^2)_{kj}$ の値は $I(i)$ 増加する．これにより，$I(i) > 1$ の場合 $(A^2)_{ij}$ の増分の方が大きくなり，$I(i)$ に応じた優先的選択よりも $(A^2)_{ij}$ に応じた優先的選択の方が，$I(j)$ の大きい j が選ばれやすくなるため，α_{ji} が小さくなりやすいのではないかと考えられる．

6　結　　論

本稿では，Twitter の日本人ユーザ 27,500,000 人から成るネットワークについての解析を行い，2-hop 先のノードを参照した優先的選択を行う成長ネットワークモデルを提案した．実データの解析から，現在のネットワークの次数分布は低次数ではスケールフリーネットワークの特徴であるべき乗則に従っておらず山なりの分布になっており，高次数では従来のべき乗分布の性質を示すことが分かった．また，提案モデルは複数の優先的選択を行い，高いクラスター性を目指すモデルであったが，単純なパラメータでは実データの低次数で見られる分布と高次数で見られる分布のどちらか一方しか表現できなかった．この

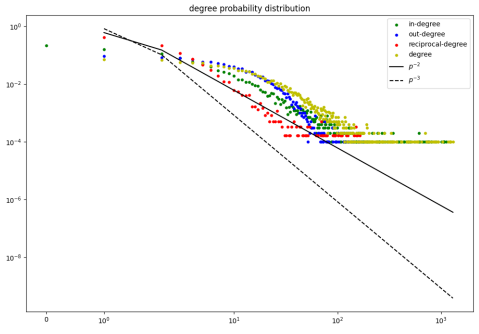

図 14　G_6 の次数分布

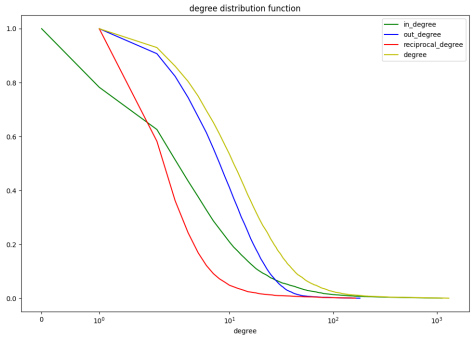

図 15　G_6 の累積分布関数

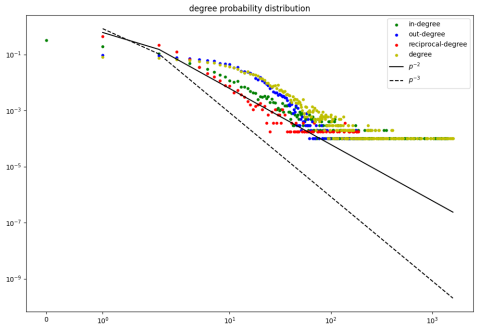

図 16　G_7 の次数分布

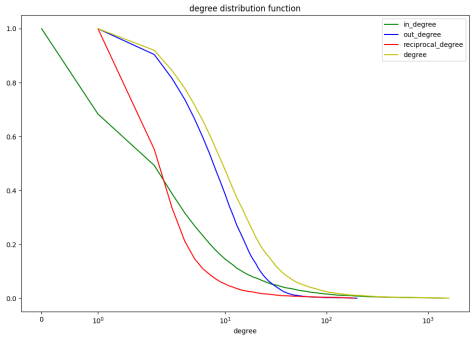

図 17　G_7 の累積分布関数

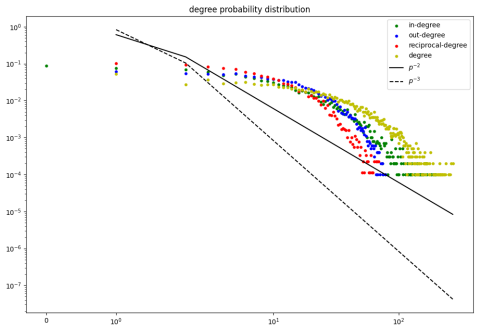

図 18　G'_1 の次数分布

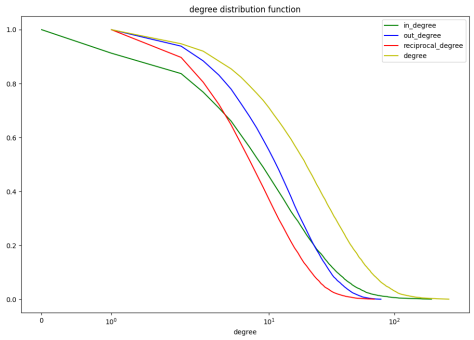

図 19　G'_1 の累積分布関数

結果から，実データではノードの次数に応じてパラメータが変化する必要があると考えられる．また，今回用いた実データはフォロワー数を上位から取得していることによるバイアスがある可能性があり，新たにフォロワー数によらないサンプリング方法を用いたデータセットを作成し同様の検証を行っていくことで，Twitter のフォローネットワークの解析を行っていきたい．

7　謝　　辞

本研究は JSPS 科研費 21H03446 の助成を受けたものです．

文　　献

[1] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, Vol. 393, No. 6684, pp. 440–442, 1998.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[3] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical review E*, Vol. 65, No. 2, p. 026107, 2002.

[4] Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, Vol. 67, No. 5, p. 056104, 2003.

[5] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, Vol. 6, No. 2, pp. 125–145, 2002.

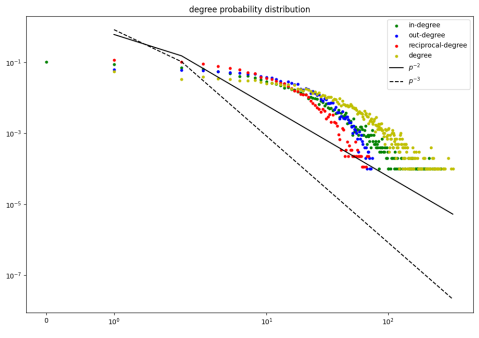

図 20　G_2' の次数分布

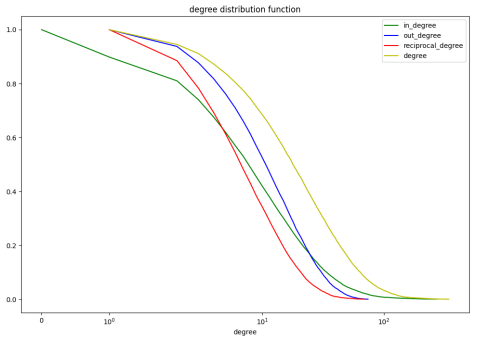

図 21　G_2' の累積分布関数

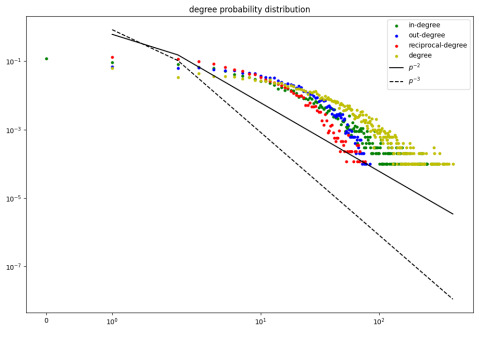

図 22　G_3' の次数分布

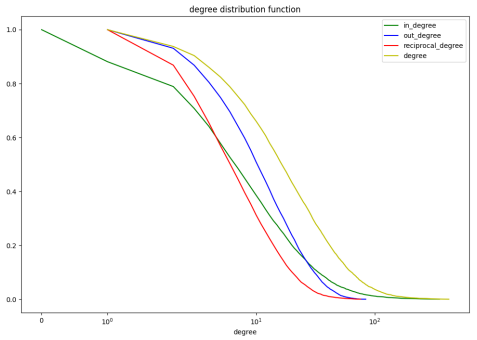

図 23　G_3' の累積分布関数

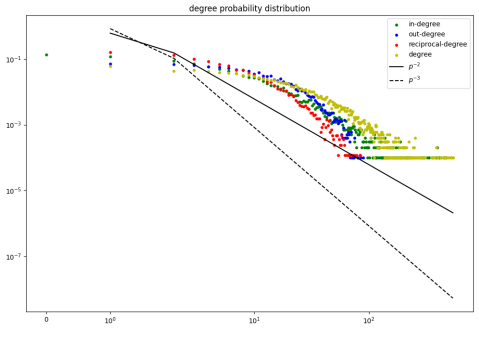

図 24　G_4' の次数分布

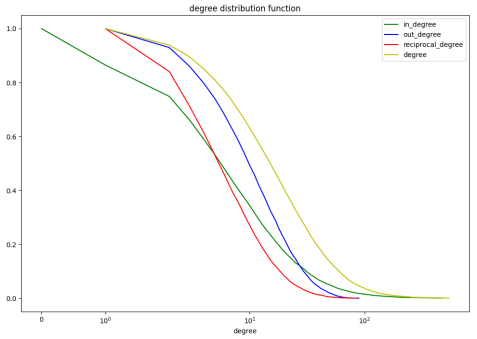

図 25　G_4' の累積分布関数

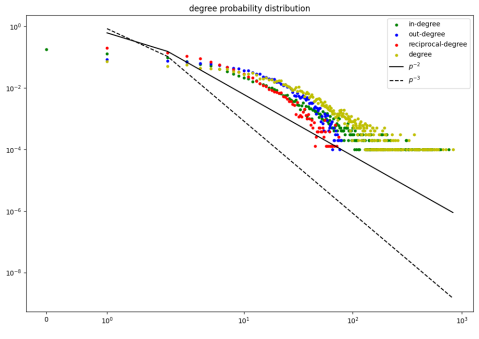

図 26　G_5' の次数分布

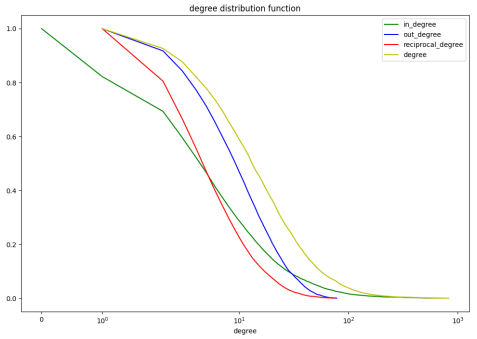

図 27　G_5' の累積分布関数

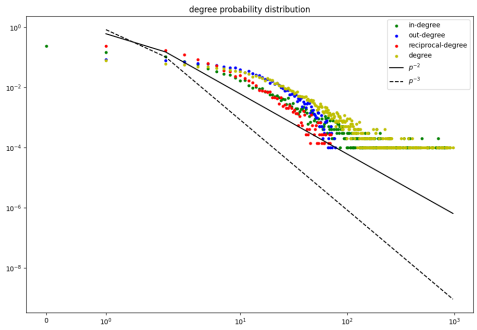

図 28　G'_6 の次数分布

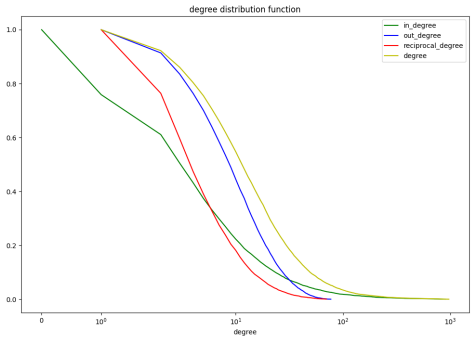

図 29　G'_6 の累積分布関数

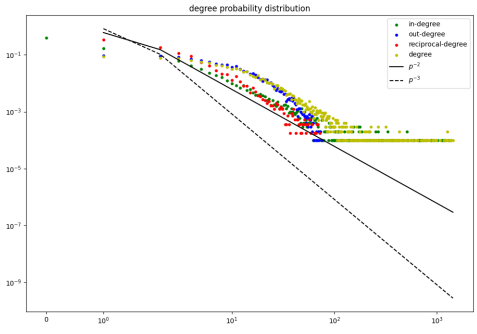

図 30　G'_7 の次数分布

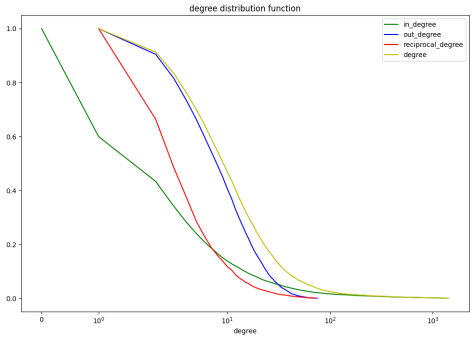

図 31　G'_7 の累積分布関数